

Ein Petrov-Galerkin-Verfahren zur
Simulation der Festkörperdiffusion
in der Halbleiterherstellung

Milos Svoboda

Disbui, München

1988





Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41554823_0068

Ein Petrov-Galerkin-Verfahren zur Simulation der Festkörperdiffusion in der Halbleiterherstellung

**Dissertation
zur Erlangung des akademischen Grads
eines Doktors der Naturwissenschaften
der Fakultät für Mathematik
der Ludwig-Maximilians-Universität, München**

**vorgelegt von
Dipl.-Math. Miloš Svoboda
aus München**

Eingereicht im Februar 1988

1. Berichterstatter:

Prof. Dr. Sachs

2. Berichterstatter:

Prof. Dr. Kalf

Tag der mündlichen Prüfung: 18. April 1988

Inhaltsverzeichnis

	Kurzfassung	1
1	Eindimensionales elliptisches Randwertproblem	4
1.1	Einleitung	4
1.2	Quadratische Polynome mit stückweise stetigen zweiten Ableitungen	4
1.3	Stückweise Konstante	6
1.4	L_2 -Approximation mit B-Splines	7
1.5	Die ASWR-Lösung für eine Raumvariable	12
2	Zweidimensionales elliptisches Randwertproblem	17
2.1	Zweidimensionale Ansatzfunktionen	17
2.2	Das zweidimensionale Randwertproblem	28
2.3	Konvergenz des ASWR-Verfahrens mit neuen Interface-Bedingungen	36
3	Zeitabhängiges Problem	40
3.1	Parabolisches Problem	40
3.2	Existenz der ASWR-Lösung	42
3.3	Konvergenz der ASWR-Lösung	43

4	Entwicklung eines Verfahrens vom ASWR-Typ für die partiellen Differentialgleichungen der Festkörperdiffusion	49
4.1	Das System partieller Differentialgleichungen zur Beschreibung der Festkörperdiffusion	49
4.2	Probleme bei der Anwendung numerischer Methoden	52
4.3	Motivation für die Untersuchung von Verfahren des ASWR-Typs für die Festkörperdiffusion	53
4.4	Transformation der Differentialgleichung	54
4.5	Semidiskrete ASWR-Lösung für die Halbleitergleichungen	55
4.6	Zeitdiskretisierung und Nichtlinearitäten	56
4.7	Orts- und zeitabhängiger Gittergenerator	56
4.8	Anpassung des Gitters für die Zeitdiskretisierung	59
4.9	Konsistente Berechnung der Lösungswerte auf einem neuen Gitter	62
4.10	Fehlerbetrachtung durch Massenerhaltung	63
5	Durchführung der ASWR-Methode für die Festkörperdiffusion	64
5.1	Kopplung von Arsen- und Bordiffusion	65
5.2	Simulation einer CMOS-Struktur	70
5.3	Ausblicke	77
	Literaturverzeichnis	78
	Bezeichnungen	81

KURZFASSUNG

Mit der fortschreitenden Entwicklung von immer höher integrierten VLSI - Bauteilen und der damit erhöhten Schwierigkeit der Erstellung immer kleinerer Strukturen bedarf es genauerer physikalischer Kenntnisse in der Modellierung dieser kleinen Strukturen. Dabei ist die genaue Berechnung der Verteilung der Störatome im Silicium ein wichtiger Punkt für die Herstellung eines Halbleiterelements.

Eine Möglichkeit der physikalischen Beschreibung der dynamischen Verteilung der Störatome ist die Transportgleichung. Dies führt unter gewissen Vereinfachungen auf ein nichtlineares parabolisches partielles Differentialgleichungssystem. Da hierfür im allgemeinen keine analytischen Lösungen bekannt sind, ist man auf die numerische Berechnung der Lösungen angewiesen. Die Simulation der physikalischen Vorgänge erfolgt hier durch die approximative Lösung eines parabolischen Differentialgleichungssystem in zwei Raumvariablen und einer Zeitvariablen. Alte Arbeiten zur numerischen Lösung benutzten finite Differenzen, was jedoch bei hohen Konzentrationen von Störatomen für die geforderte Genauigkeit zu sehr hohen Rechenzeiten führt. *Realisierungen dessen sind SUPRA[DUT-82], BICEPS[PEN-83], LADIS[TIE-80], ROMANS[MAL-83], COMPOSITE[SVO-85], und PROMIS[PIC-85]. Diese Programme basieren auf der Methode der finiten Differenzen und alle lösen mehr oder weniger eingeschränkt die in Kapitel 4 beschriebenen Gleichungen. Finite-Elemente-Programme, die obige Problematik angehen, sind derzeit entweder noch im Versuchsstadium und kämpfen mit großen Schwierigkeiten oder lösen nur vereinfachte Gleichungen, wie z.B. lineare Modelle. Diese Arbeit beschreibt in Kapitel 4 und 5 die erstmalige Anwendung des ASWR-Verfahrens in der Prozeßsimulation, wobei die ganze in Kapitel 4 vorgestellte Komplexität der Gleichungen bewältigt wird.*

Das ASWR-Verfahren (Asymetric Separated Weighted Residuals) kann als Variationsmethode betrachtet werden, ähnlich wie die Finite-Elemente-Methode. Genau genommen gehört das Verfahren zu den Petrov-Galerkin-Methoden. Bei Galerkin Methoden sind Ansatz und Testraum identisch, was bei Petrov-Galerkin-Methoden im allgemeinen nicht der Fall ist. "Asymetric" charakterisiert diese Eigenschaft. Zudem werden die zweidimensionalen Ansatzfunktionen durch separierte, eindimensionale Splinefunktionen zusammengesetzt, was eine effiziente Berechnung der diskreten Variablen ermöglicht. Daraus resultieren sogenannte "nichtkonforme" Elemente. Konvergenzaussagen für solche Verfahren sind in der Regel schwieriger zu zeigen als für Verfahren mit konformen Elementen; deswegen wird für die Anwendung bei nicht gesicherten Aussagen der sogenannte Patch-Test angewandt, mit dessen

Hilfe man die Konvergenz experimentell nachweist. Diese Arbeit erbringt eine Konvergenzaussage für ein elliptisches und parabolisches (also zeitabhängiges) Problem. Hauptergebnis sind hierbei die Konvergenzaussagen in Satz 2.3.6

$$\|u_A - u_0\|_{L_2(\Omega)} \leq c_A h^2 \|u_0\|_*$$

und Satz 3.3.4

$$\|u_0(t) - u_A(t)\|_{L_2} \leq \frac{1}{c_P} e^{-c_1(t-t_0)} \|g_A - Qg\|_P + c_A h^2 \left(\|u_0(t)\|_* + \frac{1}{c_P} \int_{t_0}^t e^{c_1(s-t)} \left\| \frac{d}{dt} u_0(s) \right\|_* ds \right).$$

Bei den Konvergenzaussagen wurde vor allem auf die Berücksichtigung einer neuartigen Interface-Bedingung (Interface-Bedingung zweiter Art) Wert gelegt. Diese Interface-Bedingung besitzt mehr noch als die Interface-Bedingung erster Art "nichtkonformen" Charakter. Das ergibt sich aus der rein physikalischen Motivation der Bedingung, die Punktstetigkeiten nicht berücksichtigt. Um eine übersichtliche Darstellung zu erreichen, wurden bei der Formulierung des Laplace-Problems konstante Koeffizienten zugrunde gelegt. Die Beweistechniken aus der Konvergenztheorie der Finiten Elemente, z.B. [ODE-76] sind auch für das ASWR-Verfahren anwendbar, siehe dazu auch [SCH-82]. Der Schwerpunkt dieser Arbeit liegt im Konvergenzbeweis für das zweidimensionale elliptische Problem unter Berücksichtigung der Interface-Bedingung zweiter Art; hierzu wird die Konstruktion einer zweidimensionalen B-Splinebasis durchgeführt, die für die praktische Anwendung von großem Vorteil ist. Diese Basisfunktionen entsprechen den Form- oder Normfunktionen bei Finiten Elementen. Sie erfüllen a priori die Interface-Bedingungen und können somit am Interface Sprünge in der Funktion darstellen. Die Grundidee des Konvergenzbeweises beruht auf der Konstruktion von "Zwischenlösungen", wie z.B. der Konstruktion einer Lösung durch Superposition von eindimensionalen ASWR-Lösungen. Es wird die Approximationsgüte dieser "Zwischenlösung" zur eigentlichen zweidimensionalen ASWR-Lösung über speziell gewählte "Filteroperatoren", welche die "Nichtkonformität" ausdrücken, qualifiziert.

Im weiteren wird in der Arbeit auf die spezifischen Schwierigkeiten im Zusammenhang mit der Implementierung des ASWR-Verfahrens für zweidimensionale nichtlineare parabolische Differentialgleichungen eingegangen werden. Dabei ergibt sich die Notwendigkeit für ein zeitabhängiges Rechengitter. In Kapitel 4 wird eine bis dahin noch nicht verwendete Transformation der Halbleitergleichung definiert. Gaussverteilungen, oft verwendete Anfangsverteilungen der Halbleitergleichungen, können mit Hilfe dieser Transformation und den

ASWR-Ansatzfunktionen mit nur einem Gitterelement ohne Approximationsfehler dargestellt werden. Zum Abschluß belegen drei numerische Beispiele die Leistungsfähigkeit des ASWR- Programms. Ein Ergebnis ist die hohe Rechengenauigkeit bei vergleichsweise für Differenzenverfahren äußerst groben Gittern. Geringe Anzahlen von diskreten Variablen führen auf niedrige Rechenzeiten selbst bei schwierigen Problemen, deren Simulation bis dahin nur unter erheblichem Aufwand möglich war. So ist eine Verminderung der Rechenzeiten um den Faktor 10 und mehr im Vergleich zu Berechnungen mit COMPOSITE [SVO-85] zu beobachten. Eine Konsequenz daraus ist die Möglichkeit der Verwendung des ASWR- Verfahrens für dreidimensionale Geometrien mit einem noch vertretbaren Rechenaufwand.

An dieser Stelle möchte ich mich herzlich bei Herrn Prof. Dr. Sachs für seine sachliche und unermüdliche Unterstützung bedanken. Herrn Prof. Dr. Ryssel, Mitinitiator für Forschungsvorhaben in zweidimensionaler Prozeßsimulation verdanke ich die Möglichkeit einer langfristigen und kontinuierlichen Mitarbeit am Programm COMPOSITE. Dipl.-Math. Dipl.-Phys. J. Lorenz stand mir dabei in kollegialer Mitarbeit zur Seite. Besonderen Dank für ihre anregende und konstruktive Kritik gebührt auch an meine Freunde Dipl.-Ing. A. Seidl, Dipl.-Math. A. Stersch und Ingrid Jusek.

1 Eindimensionales elliptisches Problem

1.1 Einleitung

Das ASWR-Verfahren (Asymmetric Separated Weighted Residual) gehört zu den Petrov-Galerkin-Methoden, deren Charakteristikum im Vergleich zu den Galerkin Methoden die nicht notwendige Gleichheit von Ansatz- und Testräumen ist. Ansatz und Testraum sind also, wie sich das durch "Asymmetric" in ASWR dokumentiert, nicht identisch. Für eine Fehler- und Konvergenzbetrachtung ist eine Abschätzung der Ansatzfunktionen Voraussetzung. Die zweidimensionalen Ansatzfunktionen (in den zwei Raumvariablen) lassen sich durch die jeweiligen eindimensionalen Ansatzfunktionen zusammensetzen; daraus resultiert "Separated" im Namen ASWR. Für eine Fehler- und Konvergenzbetrachtung ist eine Abschätzung der Ansatzfunktionen nötig. Im folgenden werden Abschätzungen für stückweise glatte Funktionen gezeigt, wobei die stückweise glatten Funktionen durch B-Splines [BOO-78] dargestellt werden. Über die Abschätzung von Quasiinterpolierenden erhält man eine Schranke für eine L_2 -Projektion in dem vom ASWR-Verfahren benötigten Testraum. Mit Hilfe der Greenschen Funktion kann die Konvergenz der eindimensionalen ASWR-Lösung gezeigt werden. Durch Superposition der jeweiligen eindimensionalen ASWR-Lösungen in den jeweiligen Raumrichtungen wird eine Lösung konstruiert, die wiederum gegen die zweidimensionale ASWR-Lösung konvergiert, womit die Konvergenz der zweidimensionalen Lösung gezeigt ist.

Die eindimensionalen Ansatzfunktionen werden durch quadratische Polynome mit stückweise stetigen zweiten Ableitungen definiert. Die Darstellung der Ansatzfunktionen erfolgt zunächst durch stückweise Polynome und später als B-Splines. Ein Satz von Curry und Schönberg belegt die Identität der beiden Darstellungen. Zunächst wird die Definition des stückweise quadratischen Polynoms mit stetigem Funktionsverlauf und stetiger Ableitung auf dem Definitionsbereich angegeben.

1.2 Quadratische Polynome mit stückweise stetigen zweiten Ableitungen

Es soll der Raum der quadratischen Polynomsplines definiert werden: Gegeben sei das abgeschlossene Intervall $[a, b]$. Weiter existiere eine streng monotone Folge $\mathbf{x} = \{x_{k-1/2}\}_1^{n+1}$ mit $x_{1/2} = a$ und $x_{n+1/2} = b$. Die Folge $\mathbf{x}$ beschreibt die Sprungstellen der zweiten Ableitungen des Polynomsplines. Der Spline selber und dessen Ableitung sollen jedoch an dieser Stelle stetig sein. Die Folge $\bar{\mathbf{x}} = \{x_k\}_1^n$ besteht aus den Mittelpunkten zwischen $x_{k-1/2}$ und $x_{k+1/2}$. $\{\mathbf{x}\}$ bezeichne die Menge aller dieser Folgen $\mathbf{x}$ mit obigen Eigenschaften. Im folgenden die Definition für den Polynomspline aus $\mathbf{P}_{2,\mathbf{x}}$:

1.2.1 Definition

Der quadratische Polynomspline p aus $\mathbf{P}_{2,\mathbf{x}}$ hat folgende Eigenschaften:

- (1) n gibt die Anzahl der Polynomstücke P_k an
- (2) die Folge $\mathbf{x}$ gibt die Intervalle für den Definitionsbereich der Polynomstücke an
- (3) p ist auf dem Intervall $[a,b]$ stetig differenzierbar.

Das Polynomstück P_k läßt sich durch die Funktionswerte an den Stellen $x_{k-1/2}$, x_k und $x_{k+1/2}$ darstellen:

1.2.2 Darstellung der Polynomsplines aus Polynomstücken

Die Polynomstücke lassen sich durch die jeweiligen Randpunkte und den Mittelpunkt des Intervalls darstellen:

1.2.3 Definition

$$P_k(x) := U_k + \frac{(x - x_k)}{h_k} (U_{k+1/2} - U_{k-1/2}) + \frac{(x - x_k)^2}{h_k^2} 2(U_{k+1/2} + U_{k-1/2} - 2U_k),$$

wobei $U_k = f(x_k)$, $U_{k-1/2} = f(x_{k-1/2})$ und $U_{k+1/2} = f(x_{k+1/2})$ ist.

Die Polynomsplines sind eindeutig durch die n Mittelpunktwerte U_k bestimmt sowie durch die zwei Randwerte $U_{1/2}$ und $U_{n+1/2}$. Dazu das nächste Lemma

1.2.4 Lemma

$p \in \mathbf{P}_{2,\mathbf{x}}$ hat die Dimension $n+2$ (Anzahl der frei wählbaren Parameter) und ist somit durch die Mittelpunktwerte und die zwei Randwerte eindeutig bestimmt.

Beweis:

Da p stetig differenzierbar ist, muß gelten:

- (1) $P_k(x_{k+1/2}) = P_{k+1}(x_{k+1/2})$ und
- (2) $\frac{\partial}{\partial x} P_k(x_{k+1/2}) = \frac{\partial}{\partial x} P_{k+1}(x_{k+1/2})$.

Wertet man (1) und (2) aus, so erhält man als Interfacebedingung:

$$(3) \quad \frac{1}{h_k} U_{k-1/2} + \left(\frac{3}{h_k} + \frac{3}{h_{k+1}}\right) U_{k+1/2} + \frac{1}{h_{k+1}} U_{k+3/2} = \frac{4}{h_k} U_k + \frac{4}{h_{k+1}} U_{k+1}.$$

Man erhält auf diese Weise $n-1$ Gleichungen. Setzt man voraus, daß die rechte Seite, $U_{1/2}$ und $U_{n+1/2}$ bekannt sind, so erhält man ein Gleichungssystem mit $n-1$ Unbekannten. Die Matrix des Gleichungssystems ist diagonaldominant und somit eindeutig lösbar.



Im folgenden benutzen wir eine spezielle B-Spline-Darstellung; sie liefert eine numerisch günstige Basis für die Berechnung von quadratischen Polynomsplines und ermöglicht eine Fehlerdarstellung, vgl. [BOO-78].

1.2.5 Definition

Sei $\mathbf{t} = \{t_i\}_1^{m+3}$ eine monoton steigende Folge auf dem Intervall $[a, b]$ mit $t_1 = t_2 = t_3 = a$, $t_{m+3} = t_{m+2} = t_{m+1} = b$ und $t_3 < t_4 < \dots < t_m < t_{m+1}$. Weiter sei $\{\mathbf{t}\}$ die Menge aller dieser Folgen $\mathbf{t}$ auf dem Intervall $[a, b]$.

Die quadratischen B-Splines zur Folge $\mathbf{t}$ haben folgendes Aussehen:

$$\begin{aligned}
 & \frac{(x - t_i)^2}{(t_{i+2} - t_i)(t_{i+1} - t_i)} && \text{für } x \in [t_i, t_{i+1}) \\
 B_{\mathbf{t},i}(x) := & \frac{(x - t_i)(t_{i+2} - x)}{(t_{i+2} - t_i)(t_{i+2} - t_{i+1})} + \frac{(t_{i+3} - x)(x - t_{i+1})}{(t_{i+3} - t_{i+1})(t_{i+2} - t_i)} && \text{für } x \in [t_{i+1}, t_{i+2}) \\
 & \frac{(t_{i+3} - x)^2}{(t_{i+3} - t_{i+2})(t_{i+3} - t_{i+1})} && \text{für } x \in [t_{i+2}, t_{i+3}) \\
 & \text{und für } i \in \{1, \dots, m\}.
 \end{aligned}$$

$$\mathbf{S}_{2,\mathbf{t}} := \text{span} (B_{\mathbf{t},1}(x), \dots, B_{\mathbf{t},m}(x)).$$

Die Dimension des von den B-Splines aufgespannten Raums $\mathbf{S}_{2,\mathbf{t}}$ ist folglich m . Ein Satz von Curry und Schönberg [BOO-79] liefert nun, daß die Räume $\mathbf{S}_{2,\mathbf{t}}$ und $\mathbf{P}_{2,\mathbf{x}}$ identisch sind, falls $\mathbf{x} \subset \mathbf{t}$ und $n+2 = m$ ist. Das leuchtet ein, wenn man berücksichtigt, daß ein $f \in \mathbf{P}_{2,\mathbf{x}}$ durch die n Mittelpunktwerte an den Stellen x_k und die beiden Randwerten an a und b eindeutig bestimmt wird.

Damit sind die Ansatzfunktionen entweder durch Polynomsplines oder B-Splines darstellbar. Eine sehr einfache Darstellung läßt sich naturgemäß für die stückweise stetigen Konstanten angeben: sie stellen die Testfunktionen für das Variationsprinzip dar.

1.3 Stückweise Konstante

Die stückweisen Konstanten werden auf der Folge $\mathbf{x}$ definiert. In den weiteren Darstellungen soll nun stillschweigend angenommen werden, daß $\mathbf{x} \subset \mathbf{t}$ und $n+2=m$. $\mathbf{t}$ unterscheidet sich also nur durch die "Vielfachheiten" am Rand von $\mathbf{x}$.

1.3.1 Definition

Die stückweise konstanten B-Splines zur Folge $\mathbf{x}$ haben folgendes Aussehen:

$$K_{\mathbf{x},k}(x) := \begin{cases} 1 & x \in [x_{k-1/2}, x_{k+1/2}), \\ 0 & \text{sonst.} \end{cases}$$

$$\mathbf{S}_{0,\mathbf{x}} := \text{span} (K_{\mathbf{x},1}(x), \dots, K_{\mathbf{x},n}(x)).$$

1.3.2 Bemerkung

Die Dimension von $\mathbf{S}_{0,\mathbf{x}}$ ist n . Dies ist sofort klar, da die n Konstanten durch n Werte eindeutig bestimmt werden.

Große Bedeutung erhält die B-Spline-Darstellung bei der praktischen Verwendung des ASWR-Verfahrens; für das ASWR spielten sie dabei die gleiche Rolle wie die Form- oder auch Normfunktionen bei den Finiten Elementen [SCH -80].

1.4 L_2 -Approximation mit B-Splines

Für das Variationsproblem ist eine Abschätzung für die Least-Squares Projektion in den $\mathbf{S}_{2,\mathbf{t}}$ und $\mathbf{S}_{0,\mathbf{x}}$ nötig. Zuerst soll die B-Spline-Theorie für die Abschätzung der Quasiinterpolierenden verwendet werden.

1.4.1 Definition

Zur Folge $\mathbf{t}$, genannt Partition auf dem Intervall $[a,b]$, werden Funktionale β_i und γ_i folgendermaßen definiert:

$$\beta_i f := f(x_i) + \frac{1}{2} (t_{i+2} - x_i)(t_{i+1} - x_i) f(x_i) \quad \text{mit } x_i = \frac{(t_{i+2} + t_{i+1})}{2} \quad \text{und}$$

$$\gamma_i f := f(x_i).$$

Die Quasiinterpolierende von f mit quadratischen Polynomsplines ist dann

$$Q_2 f := \sum_{i=1}^m (\beta_i f) B_{\mathbf{t},i}$$

und bei stückweise konstanten Funktionen

$$Q_0 f := \sum_{i=1}^m (\gamma_i f) K_{\mathbf{t},i}.$$

Carl de Boor zeigt in [BOO-78] für das Funktional β bzw. γ :

1.4.2 Lemma

Für $i, j \in \{1, \dots, m\}$ bzw. $k, l \in \{1, \dots, n\}$ gilt

$$\beta_i B_{\mathbf{t},j}(x_i) = \delta_{ij} \quad \text{bzw.} \quad \gamma_k K_{\mathbf{x},l}(x_k) = \delta_{kl}.$$

Beweis:

Der Beweis ist z.B. in [BOO-78] zu finden. □

Die "normalisierten" B-Splines sind auf dem Träger der Länge 1 normierte B-Splines.

1.4.3 Definition

$$B_{\mathbf{t},i}^{\mathbf{p}}(x) := \sqrt[p]{\frac{3}{t_{i+3} - t_i}} B_{\mathbf{t},i}(x) \quad \text{bzw.}$$

$$K_{\mathbf{x},k}^{\mathbf{p}}(x) := \sqrt[p]{\frac{1}{h_k}} K_{\mathbf{x},k}(x).$$

Das führt zum nächsten Satz.

1.4.4 Satz

Für $p \in [1, \infty]$ $\exists C_p > 1$, so daß für

$\forall \mathbf{t} \in \{\mathbf{t}\}$ b.z.w. $\mathbf{x} \in \{\mathbf{x}\}$ und

$\beta = (\beta_1, \dots, \beta_m) \in \mathbb{R}^m$ b.z.w. $\gamma = (\gamma_1, \dots, \gamma_n) \in \mathbb{R}^n$ gilt:

$$\frac{1}{C_p} \|\beta\|_p \leq \left\| \sum_{i=1}^m \beta_i B_{\mathbf{t},i}^{\mathbf{p}} \right\|_{L_p} \leq \|\beta\|_p \quad \text{b.z.w.}$$

$$\frac{1}{C_p} \|\gamma\|_p \leq \left\| \sum_{k=1}^n \gamma_k K_{\mathbf{x},k}^{\mathbf{p}} \right\|_{L_p} \leq \|\gamma\|_p.$$

Beweis:

Der Beweis wird stellvertretend für stückweise Konstante mit quadratischen B-Splines durchgeführt.

$$\left\| \sum_{i=1}^m \beta_i B_{\mathbf{t},i}^{\mathbf{p}} \right\|_{L_p}^p = \left(\frac{1}{p} + \frac{1}{q} = 1 \right)$$

$$\int_b^a \left(\sum_{i=1}^m \beta_i \sqrt[p]{\frac{3}{t_{i+3} - t_i}} B_{\mathbf{t},i}(x) \right)^p dx \geq \quad \text{Höldersche Ungleichung}$$

$$\int_b^a \left(\sum_{i=1}^m |\beta_i|^p \frac{3}{t_{i+3} - t_i} \right) B_{\mathbf{t},i}(x) \left(\sum_{i=1}^m B_{\mathbf{t},i}(x) \right)^{\frac{p}{q}} dx \leq \sum_{i=1}^m |\beta_i|^p.$$

Linke Abschätzung:

$$g := \sum_{i=1}^m \beta_i B_{\mathbf{t},i}^{\mathbf{p}} = \sum_{i=1}^m \gamma_i B_{\mathbf{t},i} \quad \text{wobei} \quad \gamma_i := \sqrt[p]{\frac{3}{t_{i+3} - t_i}}.$$

Nun ist

$$(1) \quad |\gamma_i| \leq \sum_{j=2} \left| \left(\frac{(t_{i+2} - x_i)(t_{i+1} - x_i)}{2} \right)^{(2-j)} \right| |g^{(j)}| \quad \text{wegen Lemma 1.4.2.}$$

Sei I das größte Intervall aus $\mathbf{t}$ und $x_i \in I$.

Dann gibt es auch eine Konstante c_j , für die gilt:

$$\left| \left(\frac{(t_{i+2} - x_i)(t_{i+1} - x_i)}{2} \right)^{(2-j)} \right| \leq c_j |I|^j.$$

Weiter existiert auch ein $c_{j,p}$ unabhängig von g und $\mathbf{t}$ mit

$$(2) |g^{(j)}(x_i)| \leq c_{j,p} I^{-j-\frac{1}{p}} \sqrt[p]{\int_I |g(x)|^p dx}, \quad \text{da } g|_I \text{ ein quadratisches Polynom ist.}$$

Faßt man (1) und (2) zusammen, so erhält man

$$|\beta_i|^p = |\gamma_i|^p \frac{t_{i+3} - t_i}{3} \leq |\gamma_i|^p |I| \leq \sum_{j=3}^i (c_j c_{j,p})^p \int_I |g|^p \leq c_{p,i} \int_{t_i}^{t_{i+3}} |g|^p.$$

Summiert man über alle drei Trägerintervalle auf, so erhält man

$$C_p = \sqrt[3]{3} \sum_{k=1}^{i+3} c_{p,k}.$$

□

Für die Fehlerabschätzung der Quasiinterpolierenden benutzt C. de Boor [BOO-78] den Stetigkeitsmodul:

1.4.5 Definition

Der Stetigkeitsmodul zu einer Funktion f ist wie folgt definiert:

$$\omega(f;h) := \max \{ |f(x) - f(y)| : |x - y| \leq h \ \forall \ x, y \in [a,b] \}.$$

Einsichtig aus der Definition ist folgendes Lemma:

1.4.6 Lemma

Für $f \in \mathbf{C}^1[a,b]$ gilt: $\omega(f;h) \leq h \|f'\|_{L_\infty}$.

Damit zur Existenz und Abschätzung der Quasiinterpolierenden:

1.4.7 Satz

Für alle Folgen $\mathbf{t} \subset [a,b]$ bzw. $\mathbf{x} \subset [a,b]$ aus Definition 1.2.5 existiert

$$Q : L_\infty[a,b] \longrightarrow \mathbf{S}_{2,\mathbf{t}} \text{ bzw. } \mathbf{S}_{0,\mathbf{x}} \text{ sodaß}$$

Qf die Quasiinterpolierende zu f ist. Dabei gilt für $f \in \mathbf{C}[a,b]$:

$$\|Qf - f\|_{L_\infty[a,b]} \leq c \omega(f;h).$$

Beweis:

Der Beweis soll wieder stellvertretend nur für $Qf := Q_2 f \in \mathbf{S}_{2,t}$ geführt werden. Sei $\|\cdot\|_i := \|\cdot\|_{L^\infty[t_i, t_{i+1}]}$ und $\mathbf{S}_i := \text{span} \{ B_j : j = 1, \dots, m : i \neq j \}$. $B_j := B_{t,j}$. Nach einem Korollar des Satzes von Hahn-Banach existiert ein lineares Funktional

$$\beta_i B_j = \delta_{ij}, \quad \text{wobei}$$

(1) $|\beta_i f| \leq c \|f\|_i$ gilt, falls

$$\begin{aligned} \text{dist}(B_i, \mathbf{S}_i) &= \inf \left(\left\| B_i - \sum_{\substack{j=1 \\ i \neq j}}^m b_j B_j \right\|_i \right) \\ &= \inf \left(\left\| \sum_{j=1}^m b_j B_j \right\|_i, b_i = 1 \right) \geq \frac{1}{c}. \end{aligned}$$

Nach dem letzten Satz ist das jedoch erfüllt.

Nun ist weiter $Qf = \sum_{j=1}^m (\beta_j f) B_j$.

Q ist damit ein stetiger, linearer Projektor auf $\mathbf{S}_{2,t}$. Aus dem letzten Satz geht weiter hervor, daß $\|Qf\|_i - |\beta_i f| \geq c \|f\|_i$ ist. Für ein $x \in [t_i, t_{i+1}]$ gilt:

$$\begin{aligned} |f(x) - Qf(x)| &= |f(x) - \sum_{j=i-2}^i (\beta_j f) B_j(x)| \\ &= \left| \sum_{j=i-2}^i (B_j(x) f(x) - \beta_j f B_j(x)) \right|, \quad \text{da } \sum_{j=i-2}^i B_j(x) = 1 \text{ ist.} \end{aligned}$$

Da aber auch $Qf = \sum_{j=1}^m \beta_j f B_j(x) = 1$ ist, mit $f(x) = 1$, $x \in [t_i, t_{i+1}]$ folgt für

$$\begin{aligned} |f(x) - Qf(x)| &= \left| \sum_{j=i-2}^i \beta_j (f(x) - f) B_j(x) \right| \\ &\leq \max \{ \beta_j (f(x) - f) \mid i-2 \leq j \leq i \} \\ &\leq c \max \{ \|f(x) - f\|_j \mid i-2 \leq j \leq i \} \quad \text{wegen (1)} \\ &= c \max \{ |f(x) - f(t)| \mid t_{i-1} \leq t \leq t_{i+2} \} \\ &\leq c \omega(f; h). \end{aligned}$$



Damit kommt man nun zum Kern des Absatzes:

1.4.8 Definition

Sei P der lineare Projektor in den Raum der stückweise konstanten Funktionen, also $\mathbf{S}_{0,\mathbf{x}}$ d. h.

$$P : L_2[a, b] \longrightarrow \mathbf{S}_{0,\mathbf{x}}[a, b],$$

$$f \longmapsto Pf, \quad \text{mit}$$

$$(1) \quad \int_a^b (f - Pf) g \, dx = 0 \quad \forall g \in \mathbf{S}_{0,\mathbf{x}}[a, b]$$

Pf wird die L_2 -Approximation von f in $\mathbf{S}_{0,\mathbf{x}}$ genannt.

Mit der Gram'schen Matrix läßt sich P abschätzen.

1.4.9 Definition

Die Gram'sche Matrix G sei bestimmt durch

$$G_{ij} := \left(\int_a^b \sqrt{\frac{1}{x_{i+1/2} - x_{i-1/2}}} K_i \sqrt{\frac{1}{x_{j+1/2} - x_{j-1/2}}} K_j dx \right)_{ij}$$

Wegen $K_i \in \mathbf{S}_{0,\mathbf{x}}$ ist $G_{ij} = \delta_{ik}$.

1.4.10 Satz

Für $f \in L_\infty[a, b]$ gilt:

$$\|Pf\|_{L_\infty[a, b]} \leq \sqrt{\frac{\max (x_{i+1/2} - x_{i-1/2})}{\min (x_{i+1/2} - x_{i-1/2})}} \|f\|_\infty.$$

Beweis:

Da $Pf \in \mathbf{S}_{0,\mathbf{x}}$ ist, existiert auch eine B-Spline Darstellung von Pf :

$$Pf = \sum_{i=1}^n \gamma_i \bar{K}_i, \quad \text{wobei } \bar{K}_j := K_{\mathbf{x},j}^2 \text{ ist.}$$

Da die K_i 's nur Werte 1 und 0 annehmen können, gilt die Abschätzung:

$$(I) \quad \|Pf\|_{L_\infty[a, b]} \leq \|\gamma\|_\infty \sqrt{\max (x_{i+1/2} - x_{i-1/2})}.$$

$$\text{Nun sei } \alpha_i := \int_a^b \bar{K}_i f dx \text{ und es gilt } \gamma_i = G^{-1} \alpha_i.$$

Also folgt für (I):

$$\|Pf\|_{L_\infty[a, b]} \leq \|G^{-1}\|_\infty \|\alpha\|_\infty \sqrt{\max (x_{i+1/2} - x_{i-1/2})}.$$

Es gilt: $|\alpha_i| \leq \|f\|_\infty \int_a^b \bar{K}_i dx = \|f\|_\infty \sqrt{(x_{i+1/2} - x_{i-1/2})}$. Damit folgt

$$\|Pf\|_{L_\infty[a, b]} \leq \sqrt{\frac{\max (x_{i+1/2} - x_{i-1/2})}{\min (x_{i+1/2} - x_{i-1/2})}} \|f\|_\infty.$$



Den Abschluß bildet nun der nächste Satz, der die punktweise Konvergenz der L_2 -Approximation zeigt.

1.4.11 Satz

Sei $f \in \mathbf{C}^1[a,b]$ und h die Länge des größten Intervalls auf $[a,b]$. Dann gilt:

$$\|f - Pf\|_{L_\infty[a,b]} \leq c \left(1 + \max_{i,j=1,n} \left(\sqrt{\frac{(x_{i+1/2} - x_{i-1/2})}{(x_{j+1/2} - x_{j-1/2})}} \right) \right) h \|f'\|_{L_\infty[a,b]}.$$

Beweis:

Es existiert eine Quasiinterpolierende von f aus $\mathbf{S}_{0,\mathbf{x}}$ mit der Abschätzung aus Satz 1.4.7 $\|Qf - f\|_{L_\infty[a,b]} \leq c \omega(f;h)$.

Da $Qf \in \mathbf{S}_{0,\mathbf{x}}$ ist, gilt $Qf = PQf$. Also folgt:

$$\begin{aligned} \|f - Pf\|_{L_\infty[a,b]} &\leq \|f - Qf\|_{L_\infty[a,b]} + \|PQf - Pf\|_{L_\infty[a,b]} \\ &\leq (1 + \|P\|_{L_\infty[a,b]}) \|f - Qf\|_{L_\infty[a,b]} \end{aligned}$$

Mit 1.4.6 und 1.4.7 folgt die Behauptung. $\square$

Nun soll mit Hilfe der Greenschen Funktion die eindimensionale ASWR-Lösung konstruiert werden.

1.5 Die ASWR-Lösung für eine Raumvariable

Das ASWR-Verfahren gehört zu den Petrov-Galerkin-Methoden: es läßt sich somit als Projektionsverfahren (Weighted Residuals) formulieren. Es soll der Einfachheit halber ein lineares Laplace-Problem mit konstanten Koeffizienten und Dirichlet-Randbedingungen betrachtet werden.

1.5.1 Definition

Das Randwertproblem habe folgende Form:

$$\begin{aligned} L u(x) &= r(x) & x \in [a,b], \text{ wobei} \\ L : W_{2,\infty}[a,b] &\longrightarrow L_\infty[a,b], \\ u &\longmapsto k u_{xx}, \quad k \text{ konstant,} \end{aligned}$$

mit den Randbedingungen:

$$u(a) := r_l, \quad u(b) := r_r, \quad r_l, r_r \in \mathbb{R}.$$

Damit kann die ASWR-Lösung zum Randwertproblem definiert werden:

1.5.2 Definition

$u_A \in \mathbf{S}_{2,t}$ ist ASWR-Lösung zum Randwertproblem genau dann, falls

$$(Lu_A - r, v)_{L_2[a,b]} = 0 \quad \forall v \in \mathbf{S}_{0,x} \quad \text{und}$$

u_A die Randbedingungen erfüllt.

Wie schon erwähnt, sind Ansatz- und Testraum nicht identisch. Somit wird im Gegensatz zu den Finiten-Elementen nicht mit den schwachen, sondern mit den starken Lösungen gearbeitet. Auf den Laplace-Operator wird nicht der Gauss'sche Integralsatz angewandt, um damit die partiellen Ableitungen auf Test- und Ansatzraum zu verteilen. Beim ASWR wird die Symmetrie erst durch die einseitige Anwendung des Laplace-Operators auf den Ansatzraum erreicht. Wenn $u_A \in \mathbf{S}_{2,t}$ ist, so ist entsprechend $Lu_A \in \mathbf{S}_{0,x}$ und somit (Lu_A, v) symmetrisch.

Um nun eine ASWR-Lösung aus der tatsächlichen Lösung zu konstruieren, verwendet man die Greensche Funktion:

1.5.3 Definition

Sei Q_1 das Dreieck $a \leq s \leq x \leq b$,
 Q_2 das Dreieck $a \leq x \leq s \leq b$ und
 $Q_g = Q_1 \cup Q_2$.

Für die Grundleistung $g(x,s)$ zum Randwertproblem gelte:

$g(x,s)$ ist stetig auf Q_g .

g_x und g_{xx} existieren in Q_1 und Q_2 und sind bis auf die Diagonale stetig.

für festes $s \in [a,b]$ ist $g(x,s)$ Lösung von $Lg(\cdot, s) = 0$ für $x \neq s$,

auf der Diagonalen macht die erste Ableitung einen Sprung der Größe k .

$\Gamma(x,s)$ ist Greensche Funktion zum Randwertproblem, falls

$\Gamma(x,s)$ Grundleistung ist und

$\Gamma(x,s)$ die Randbedingungen erfüllt.

Damit zur Definition des Greenschen Operators:

$$G : L_\infty[a,b] \longrightarrow W_{2,\infty}[a,b] \quad \text{mit}$$

$$r \longmapsto Gr := \int_a^b \Gamma(\cdot, s) r(s) ds.$$

Mit Gr ist die eindeutig bestimmte Lösung des Randwertproblems konstruierbar. Mit dieser Technik läßt sich auch die Existenz einer eindeutig bestimmten ASWR-Lösung nachweisen.

1.5.4 Satz

Ist $u_0 \in C^2[a, b]$ die eindeutig bestimmte Lösung des Randwertproblems, so existiert eine eindeutig bestimmte ASWR-Lösung $u_A \in S_{2,t}[a, b]$.

Beweis:

Aufgrund des gegebenen Randwertproblems gilt durch $u_{xx} = \frac{r}{k}$

$$(1) \quad Gu_{xx} = u.$$

Aus der Definition der ASWR-Lösung folgt mit $k \neq 0$:

$$(u_{A_{xx}}, v)_{L_2[a, b]} = \left(\frac{r}{k}, v\right)_{L_2[a, b]} \quad \forall v \in S_{0,x}[a, b]$$

$$\langle \Rightarrow \rangle \quad Pu_{A_{xx}} = P \frac{r}{k}$$

Die rechte Seite existiert, somit auch die linke. Mit (1) ist die Existenz gesichert. u_A läßt sich folgendermaßen konstruieren:

$$(2) \quad u_A = G P \frac{1}{k} L u_0.$$

Zum Abschluß des Kapitels die Konvergenz für das eindimensionale Randwertproblem:

1.5.5 Satz

Sei $u_0 \in C^3[a, b]$ die eindeutig bestimmte Lösung des Randwertproblems. Dann gelten für die ASWR-Lösung $u_A \in S_{2,t}[a, b]$ folgende Abschätzungen:

$$(1) \quad \|u_A'' - u_0''\|_{L_\infty[a, b]} \leq c_2 h \|u_0^{(3)}\|_{L_\infty[a, b]},$$

$$(2) \quad \|u_A' - u_0'\|_{L_\infty[a, b]} \leq c_1 h^2 \|u_0^{(3)}\|_{L_\infty[a, b]},$$

$$(3) \quad \|u_A - u_0\|_{L_\infty[a, b]} \leq c_2 h^2 \|u_0^{(3)}\|_{L_\infty[a, b]}.$$

h ist dabei die Länge des größten Intervalls aus t bzw. x .

Beweis:

Nach Satz 1.4.11 gilt:

$$\|Pu_0'' - u_0''\|_{L_\infty[a, b]} \leq c_a h \|u_0^{(3)}\|_{L_\infty[a, b]},$$

$$\text{weiter ist} \quad u_0'' = \frac{r}{k} \quad \text{und} \quad u_A'' = P \frac{r}{k}.$$

Daraus folgt

$$(4) \quad \|u_A'' - u_0''\|_{L_\infty[a, b]} = \left\| \frac{r}{k} - P \frac{r}{k} \right\|_{L_\infty[a, b]} = \|Pu_0'' - u_0''\|_{L_\infty[a, b]}$$

und damit (1).

Ferner gilt:

$$\begin{aligned} \|u_A' - u_0'\|_{L_\infty[a, b]} &= \|G'(u_A'' - u_0'')\|_{L_\infty[a, b]} \\ &\leq \|G'\|_\infty \|u_A'' - u_0''\|_{L_\infty[a, b]} \\ &= \|G'\|_\infty \|u_A'' - P u_A''\|_{L_\infty[a, b]}. \end{aligned}$$

Dabei ist G' :

$$\begin{array}{ccc} G' : L_\infty[a,b] & \longrightarrow & W_{1,\infty}[a,b] \quad \text{mit} \\ r & \longmapsto & G'r := \int_a^b \frac{\partial}{\partial x} \Gamma(\cdot, s) r(s) ds. \end{array}$$

Analog ist

$$\begin{aligned} \|u_A - u_0\|_{L_\infty[a,b]} &= \|G(u_A'' - u_0'')\|_{L_\infty[a,b]} \\ &\leq \|G\|_\infty \|u_A'' - u_0''\|_{L_\infty[a,b]} \\ &\leq \|G\|_\infty \|u_A'' - Pu_A''\|_{L_\infty[a,b]}. \end{aligned}$$

Da $Lu_0 = r$ ist, gilt nach der Definition der ASWR-Lösung:

$$(5) \quad \int_b^a L(u_A - u_0) K_{i,j}(s) ds = 0 \quad \forall i = 1, \dots, n$$

Sei nun $e := u_A - u_0$, dann ist

$$(6) \quad \int_{x_{i-1/2}}^{x_{i+1/2}} Le(s) ds = 0 \quad \forall i = 1, \dots, n.$$

Sei $\Gamma(y, z)$ Greensche Funktion zum Randwertproblem. Da u_A und u_0 die Randbedingungen erfüllen, so gilt dies auch für e . Somit gilt:

$$\begin{aligned} e'(x) &= \int_a^b \frac{\partial}{\partial x} \Gamma(x, s) Le(s) ds. \\ (7) \quad &= \sum_{l=1}^n \int_{x_{l-1/2}}^{x_{l+1/2}} \frac{\partial}{\partial x} \Gamma(x, s) Le(s) K_{x,l}(s) ds, \\ &\quad \text{da } K_{x,l}(x) = 1 \text{ auf } [x_{l-1/2}, x_{l+1/2}] \text{ ist.} \end{aligned}$$

Aus der Definition der Greenschen Funktion ist ersichtlich, daß man

$\frac{\partial}{\partial x} \Gamma(x, s)$ abschätzen kann. Es gilt also:

$$(8) \quad c_1 := \sup \left\{ \left| \frac{\partial}{\partial x} \Gamma(x, s) \right| \mid (x, s) \in [a, b]^2 \text{ mit } x \neq s \right\} < \infty.$$

Jetzt betrachtet man eine Abschätzung für das Intervall $[x_{i-1/2}, x_{i+1/2}]$ und das Komplementärintervall auf $[a, b]$.

Für $x \in [x_{i-1/2}, x_{i+1/2}]$ gilt mit (7) und (8):

$$(9) \quad |e'(x)| \leq c_1 (x_{i+1/2} - x_{i-1/2}) \|Le\|_{L_\infty[a,b]}.$$

Für das Komplementärintervall $[a, b] \setminus [x_{i-1/2}, x_{i+1/2}]$ benötigen wir folgendes Ergebnis aus dem Satz von Taylor:

Es seien $\varphi_1 \in \mathbf{C}^1[x_{k-1/2}, x_{k+1/2}]$ und $\varphi_2 \in \mathbf{C}[x_{k-1/2}, x_{k+1/2}]$ mit

$$\int_{x_{k-1/2}}^{x_{k+1/2}} \varphi_2(x) \, dx = 0.$$

Nach dem Satz von Taylor gilt mit $\xi(x) \in [x_{k-1/2}, x]$

$$\begin{aligned} (10) \quad & \left| \int_{x_{k-1/2}}^{x_{k+1/2}} \varphi_1(x) \varphi_2(x) \, dx \right| = \\ & = \left| \int_{x_{k-1/2}}^{x_{k+1/2}} (\varphi_1(x_{k-1/2}) + (x - x_{k-1/2}) \varphi_1'(\xi(x))) \varphi_2(x) \, dx \right| \\ & = \left| \int_{x_{k-1/2}}^{x_{k+1/2}} (x - x_{k-1/2}) \varphi_1'(\xi(x)) \varphi_2(x) \, dx \right| \\ & \leq \|\varphi_1'\|_{L_\infty[x_{k-1/2}, x_{k+1/2}]} \|\varphi_2\|_{L_\infty[x_{k-1/2}, x_{k+1/2}]} \frac{(x_{k-1/2}, x_{k+1/2})^2}{2}. \end{aligned}$$

Mit $c_2 := \sup\left\{\left|\frac{\partial}{\partial s} \frac{\partial}{\partial x} \Gamma(x, s)\right| : (x, s) \in [a, b]^2 \text{ mit } x \neq s\right\} \quad \infty$ folgt aus (5),

(6) und (10), für $x \in [a, b] \setminus [x_{i-1/2}, x_{i+1/2}]$:

$$\begin{aligned} |e'(x)| & \leq \sum_{i=1}^n c_2 \frac{1}{2} (x_{i+1/2} - x_{i-1/2})^2 \|Le\|_{L_\infty[a, b]} \\ & = \frac{1}{2} \max_{i=1, n} (x_{i+1/2} - x_{i-1/2}) c_2 \|Le\|_{L_\infty[a, b]} \quad \text{mit} \\ & \quad \max_{i=1, n} (x_{i+1/2} - x_{i-1/2}) \leq 1. \end{aligned}$$

Zusammen mit (9) ergibt sich eine Abschätzung für das ganze Intervall $[a, b]$:

$$\begin{aligned} |e'(x)| & \leq (c_1 (x_{i+1/2} - x_{i-1/2}) + \frac{1}{2} \max_{i=1, n} (x_{i+1/2} - x_{i-1/2}) c_2) \|Le\|_{L_\infty[a, b]} \\ & \leq c \, h \, \omega(u_0''; h) \\ & \leq c \, h^2 \|u_0^{(3)}\|_{L_\infty[a, b]} \quad \text{wegen Lemma 1.4.6 und damit folgt (2).} \end{aligned}$$

Analog erhält man Behauptung (3).



2 Zweidimensionales elliptisches Randwertproblem

Die Ansatzfunktionen für den zweidimensionalen Fall lassen sich, wie durch "separated" im Namen ASWR dokumentiert, durch eine Summe von eindimensionalen, in der jeweiligen Raumrichtung definierten Ansatzfunktionen zusammensetzen. Es werden "nicht-konforme" Elemente benutzt, was bedeutet, daß der Laplace-Operator an den Diskretisierungsschnittflächen wegen zu geringer Differenzierbarkeit nicht mehr angewandt werden kann. Dieser Problematik gerecht werdend, wird die Konvergenzaussage über mehrere Teilschritte erreicht. Zuerst wird durch Superposition und einen Filteroperator ein Element aus dem zweidimensionalen Ansatzraum konstruiert, das sich aus eindimensionalen Lösungen bei einer festgehaltenen Komponente des zweidimensionalen Randwertproblems ergibt. Dann wird gezeigt, daß diese Lösung gegen die eigentliche zweidimensionale ASWR-Lösung konvergiert. Weiter werden zweidimensionale Ansatzfunktionen konstruiert, deren Integrale über die Grenzflächen stetig sind, die also den physikalischen Gegebenheiten im zweidimensionalen Problem entsprechen. Durch einen weiteren "Filteroperator" läßt sich deren Konvergenz folgern. Damit kann die Konvergenz für Ansatzfunktionen gezeigt werden, die nicht notwendig am Rand stetig sind.

2.1 Zweidimensionale Ansatzfunktionen

In diesem Absatz werden zwei verschiedene Arten von Kopplungen der Boxen vorgestellt. Bei der ersten Möglichkeit der Interfacebedingung wird im Gegensatz zur zweiten Möglichkeit zumindest an speziellen Interfacepunkten die Punktstetigkeit gefordert. In beiden Fällen werden die Ansatzfunktionen im Inneren der Boxen durch die Summe von eindimensionalen quadratischen Polynomen erzeugt. In beiden Fällen lassen sich also die Ansatzfunktionen durch folgende Definition festlegen:

2.1.1 Definition

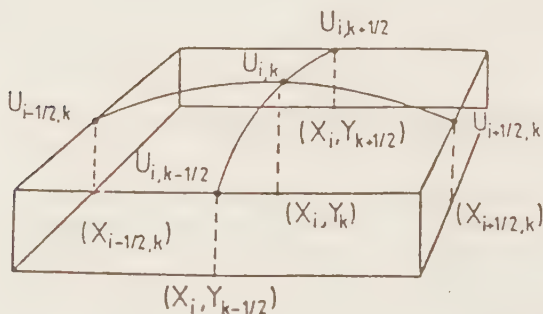
Ohne Einschränkung der Allgemeinheit sei das zweidimensionale Gebiet durch $\Omega := [a,b] \times [a,b]$ definiert. Auf diesem sei analog zum eindimensionalen Fall $\mathbb{B} := \mathbb{X} \times \mathbb{X}$ und $\mathbb{P} := \mathbb{t} \times \mathbb{t}$ mit den jeweiligen Analoga zwischen $\mathbb{X}$ und $\mathbb{t}$, also z.B. $\mathbb{P} \supset \mathbb{B}$. Innerhalb einer Box $\Omega_{i,k}$ ist die quadratische zweidimensionale Ansatzfunktion durch die Summe der unter 1.2 definierten eindimensionalen Ansatzfunktionen zusammengesetzt:

$$P_{i,k}(x,y) := P_i(x) + P_k(y) + U_{i,k} \quad \text{mit}$$

$$P_i(x) := \frac{(x - x_i)(U_{i+1/2,k} - U_{i-1/2,k})}{h_i} + \frac{(x - x_i)^2}{h_i^2} 2(U_{i+1/2,k} + U_{i-1/2,k} - 2U_{i,k})$$

$$P_k(y) := \frac{(y - y_k)(U_{i,k+1/2} - U_{i,k-1/2})}{h_k} + \frac{(y - y_k)^2}{h_k^2} 2(U_{i,k-1/2} + U_{i,k+1/2} - 2U_{i,k})$$

wobei $U_{i,k} = f(x_i, y_k)$, $U_{i,k-1/2} = f(x_i, y_{k-1/2})$, $U_{i,k+1/2} = f(x_i, y_{k+1/2})$
 $U_{i-1/2,k} = f(x_{i-1/2}, y_k)$, $U_{i+1/2,k} = f(x_{i+1/2}, y_k)$.



Damit sind die zweidimensionalen Ansatzfunktionen in der Box $\Omega_{i,k}$ definiert. Was fehlt, um den Ansatzraum vollständig zu definieren, sind die Interfacebedingungen für die Kopplung der lokalen Polynome. Dazu werden, wie schon oben erwähnt, zwei verschiedene Möglichkeiten angeboten. Damit also zu den zwei Definitionen der Ansatzfunktionen:

2.1.2 Definition

Kopplung der Boxen Ω_{ik} der ersten Art:

Stetigkeit der Polynome und der ersten Ableitungen an bestimmten Interfacebedingungen:

$$P_{ik}(x_{i+1/2}, y_k) = P_{i+1,k}(x_{i+1/2}, y_k) ,$$

$$\frac{\partial}{\partial x} P_{ik}(x_{i+1/2}, y_k) = \frac{\partial}{\partial x} P_{i+1,k}(x_{i+1/2}, y_k) .$$

Analog am x-Rand:

$$P_{ik}(x_i, y_{k+1/2}) = P_{i,k+1}(x_i, y_{k+1/2}) ,$$

$$\frac{\partial}{\partial x} P_{ik}(x_i, y_{k+1/2}) = \frac{\partial}{\partial x} P_{i,k+1}(x_i, y_{k+1/2}) .$$

2.1.3 Bemerkung

Die Stetigkeit der Ansatzfunktionen ist also am Interface nur im Punkt $(x_{i+1/2}, y_k)$ und analogen Punkten gegeben. Damit ist auch klar, daß der Laplace-Operator an diesen Stellen nicht anwendbar ist.

2.1.4 Definition

Kopplung der Boxen Ω_{ik} der zweiten Art:

Stetigkeit der Randintegrale über den Polynomen und deren Ableitungen:

$$\int_{y_{k-1/2}}^{y_{k+1/2}} P_{ik}(x_{i+1/2}, y) dy = \int_{y_{k-1/2}}^{y_{k+1/2}} P_{i+1,k}(x_{i+1/2}, y) dy ,$$

$$\int_{y_{k-1/2}}^{y_{k+1/2}} \frac{\partial}{\partial x} P_{ik}(x_{i+1/2}, y) dy = \int_{y_{k-1/2}}^{y_{k+1/2}} \frac{\partial}{\partial x} P_{i+1,k}(x_{i+1/2}, y) dy .$$

Analog für das Interface in x-Richtung

$$\int_{x_{i-1/2}}^{x_{i+1/2}} P_{ik}(x, y_{k+1/2}) dx = \int_{x_{i-1/2}}^{x_{i+1/2}} P_{i,k+1}(x, y_{k+1/2}) dx ,$$

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \frac{\partial}{\partial x} P_{ik}(x, y_{k+1/2}) dx = \int_{x_{i-1/2}}^{x_{i+1/2}} \frac{\partial}{\partial x} P_{i,k+1}(x, y_{k+1/2}) dx .$$

2.1.1 zusammengefaßt mit 2.1.2 bzw. 2.1.4 ergibt den Raum für die Ansatzfunktionen. Wie schon im eindimensionalen Fall soll dafür eine B-Spline-Repräsentation angegeben werden. Damit erhält man die für die Praxis wichtigen Formfunktionen; zugleich kann man damit auch Existenz und Eindeutigkeit einer zweidimensionalen ASWR-Lösung nachweisen. Weiter lassen sich damit auch die Ergebnisse aus Kapitel 1 anwenden. Vor der B-Spline-Darstellung jedoch noch eine Bemerkung zur Anzahl der Freiheitsgrade.

2.1.5 Bemerkung

Die Ansatzfunktionen der ersten bzw. zweiten Art sind bei Abdeckung der Randbedingungen eindeutig durch die Mittelpunktswerte bestimmt. Dies resultiert aus Lemma 1.4.2.

2.1.6 Definition

$\mathbf{S}_n$ sei der Ansatzraum für die Funktionen der ersten Art.

$\mathbf{D}_n$ sei der Ansatzraum für die Funktionen der zweiten Art.

2.1.7 Definition

$B_{ik}(x,y)|_{\Omega_{jl}} := B_i(x) B_k(y_1) + B_k(y) B_i(x_j) - B_i(x_j) B_k(y_1)$, wobei

$B_i(x)$ und $B_k(y)$ die jeweiligen eindimensionalen B-Splines in x - bzw. y -Richtung zur jeweiligen eindimensionalen Partition $\mathbf{t}$ sind.

2.1.8 Satz

$B_{ik}(x,y) \in \mathbf{S}_n$.

Beweis:

Nachprüfung der Interfacebedingungen erster Art.

Stetigkeit im Punkt $(x_{j+1/2}, y_k)$:

$$\begin{aligned} B_{ik}(x_{j+1/2}, y_1)|_{\Omega_{jl}} &= B_i(x_{j+1/2}) B_k(y_1) + B_k(y_1) B_i(x_j) - B_i(x_j) B_k(y_1) \\ &= B_i(x_{j+1/2}) B_k(y_1) \\ &= B_i(x_{j+1/2}) B_k(y_1) + B_k(y_1) B_i(x_{j+1}) - B_i(x_{j+1}) B_k(y_1) \\ &= B_{ik}(x_{j+1/2}, y_1)|_{\Omega_{j+1,1}} \end{aligned}$$

Stetigkeit der Ableitung im Punkt $(x_{j+1/2}, y_k)$:

$$\begin{aligned} \frac{\partial}{\partial x} B_{ik}(x_{j+1/2}, y_1)|_{\Omega_{jl}} &= \frac{\partial}{\partial x} B_i(x_{j+1/2}) B_k(y_1) \\ &= \frac{\partial}{\partial x} B_{ik}(x_{j+1/2}, y_1)|_{\Omega_{j+1,1}} \end{aligned}$$

Analoges gilt für die anderen Interfacepunkte.

Für den folgenden Satz ist die Definition des Tensorproduktes zweier Matrizen und ein Lemma nötig.

2.1.9 Definition

Seien $A := (a_{ij})$ und $B := (b_{kl})$ (quadratische) Matrizen der Dimension m bzw. n . Dann ist das Tensorprodukt von A und B

$$A \times B := \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & & \vdots \\ a_{n1}B & \dots & a_{nn}B \end{bmatrix} = \begin{bmatrix} d_{1111} & \dots & d_{1n1m} \\ \vdots & & \vdots \\ d_{n1m1} & \dots & d_{nnmm} \end{bmatrix}$$

eine Matrix der Dimension mn .

2.1.10 Lemma

Seien A und B wie in der obigen Definition gegeben und dazu symmetrisch und reell. Zudem seien $(\lambda_i)_{i=1,m}$ die Eigenwerte von A und $(\mu_k)_{k=1,n}$ die Eigenwerte von B . Dann hat $A \times B$ die Eigenwerte $(\lambda_i \mu_k)_{i=1,m; k=1,n}$.

Beweis:

A und B lassen sich in Diagonalform bringen, das heißt, es existieren Matrizen J und K , so daß gilt:

$$A = JD_1J^{-1} \quad \text{und} \quad B = KD_2K^{-1}, \quad \text{wobei}$$

$$D_1 = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix}, \quad D_2 = \begin{bmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_n \end{bmatrix} \quad \text{ist.}$$

Nun gelten für das Tensorprodukt folgende Rechenregeln:

- (1) $(E_1 \times F_1)(E_2 \times F_2) = E_1 E_2 \times F_1 F_2$
- (2) $E \times F = M(F \times E)M^T$, wobei $|M| = |M^T| = \pm 1$
mit $|\cdot|$ als Determinantenfunktion.
- (3) $|E \times F| = |E|^m |F|^n$, wobei E eine $m \times m$ - und F eine $n \times n$ -Matrix ist.

Es folgt (3) aus (1) da $|I_n \times F| = |F|^n$ mit $I_n \times n$ Einheitsmatrix und

$$|E \times I_m| = M(I_m \times E)M^T = |M|^2 |E|^m = |E|^m \quad \text{ist.}$$

Insbesondere gilt für quadratische und reguläre Matrizen F und E :

$$(4) \quad (F \times E)^{-1} = (F^{-1} \times E^{-1}).$$

Damit ist die Behauptung schnell gezeigt:

$$\begin{aligned} A \times B &= (J D_1 J^{-1}) \times (K D_2 K^{-1}) \\ &= (J \times K) (D_1 \times D_2) (J^{-1} \times K^{-1}) \\ &= (J \times K) (D_1 \times D_2) (J \times K)^{-1}, \text{ das hei\ss}t, A \times B \text{ kann man in Diagonalform} \\ &\quad \text{bringen, wobei} \end{aligned}$$

$$D_1 \times D_2 = \begin{bmatrix} \lambda_1 \mu_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \mu_n \end{bmatrix}$$

die Eigenwerte des Tensorproduktes darstellt.



2.1.11 Satz

$B_{ik}(x,y)$ ist Basis zu $\mathbf{S}_n$.

Beweis:

Ohne Einschränkung der Allgemeinheit sei zur einfacheren Darstellung das Gitter äquidistant gewählt; somit sind die Boxen alle gleich groß. Es reicht, aufgrund des letzten Satzes zu zeigen, daß ein Element aus $\mathbf{S}_n$ durch eine Linearkombination der B_{ik} eindeutig darstellbar ist. Es reicht sogar zu zeigen, daß die Mittelpunktswerte eindeutig darstellbar sind, da aufgrund von Bemerkung 2.1.3 die Elemente aus $\mathbf{S}_n$ eindeutig durch die Mittelpunktswerte festgelegt sind.

Sei nun $G_r := (B_i(x_j))_{ij}$. Aus der Definition 1.2.5 folgt

$$G_r = \frac{1}{8} \begin{bmatrix} 5 & 1 & & 0 \\ 1 & 6 & 1 & \\ & \ddots & \ddots & \\ 0 & & 1 & 6 & 1 \\ & & & 1 & 5 \end{bmatrix}, \quad G_r \text{ ist also positiv definit.}$$

Weiter ist $(B_{ik}(x_j, y_l))_{jl} = G_r \times G_r$, also das Tensorprodukt der Matrix G_r mit sich selbst. Da G_r positiv definit ist, gilt dies auch für $G_r \times G_r$ nach Lemma 2.1.10 und damit ist die Eindeutigkeit gezeigt.



Nun zur Basis für die Ansatzfunktionen zweiter Art. Wie oben schon erwähnt, unterscheiden diese sich durch die Integralbedingungen im Interface, was sich auch in der Darstellung der Basiselemente zeigt.

2.1.12 Definition

$$\text{Sei} \quad \bar{B}_{kl} := \frac{1}{h_{y1}} \int_{y_{l-1/2}}^{y_{l+1/2}} B_k(y) dy, \quad \bar{B}_{ij} := \frac{1}{h_{xj}} \int_{x_{j-1/2}}^{x_{j+1/2}} B_i(x) dx.$$

$$\text{Dann ist} \quad D_{ik}(x,y)|_{\Omega_{jl}} := \bar{B}_{kl} B_i(x) + \bar{B}_{ij} B_k(y) - \bar{B}_{kl} \bar{B}_{ij}.$$

(Im äquidistanten Fall ist $h_{y_k} = h_{x_i} = h$)

Wie schon erwähnt, wird i.a. an dem Interface keine Stetigkeit sein. Da die eindimensionalen B-Splines jedoch stetig sind, wird der Sprung durch die Gewichtung der Integralwerte $\bar{B}$ erzeugt. Für die Funktion $x^2 + y^2$ wird die Approximation jedoch identisch sein mit der ersten Art, denn beide Splines können die Funktion exakt darstellen. Der Vorteil der zweiten Art, um diesen noch einmal zu erwähnen, liegt in der besseren räumlichen Kopplung der einzelnen Boxen miteinander.

2.1.13 Satz

$$D_{ik}(x,y) \in \mathbf{D}_n.$$

Beweis:

Nachweis der Stetigkeit der Randintegrale:

$$\begin{aligned} & \int_{y_{l-1/2}}^{y_{l+1/2}} D_{ik}(x_{j+1/2}, y) |_{\Omega_{jl}} dy = \\ & h_{y1} \bar{B}_{kl} B_i(x_{j+1/2}) + \bar{B}_{ij} \int_{y_{l-1/2}}^{y_{l+1/2}} B_k(y) dy - h_{y1} \bar{B}_{kl} \bar{B}_{ij} = \\ & h_{y1} \bar{B}_{kl} B_i(x_{j+1/2}). \end{aligned}$$

$$\begin{aligned} & \int_{y_{l-1/2}}^{y_{l+1/2}} D_{ik}(x_{j+1/2}, y) |_{\Omega_{j+1,l}} dy = \\ & h_{y1} \bar{B}_{kl} B_i(x_{j+1/2}) + \bar{B}_{i+1,j} \int_{y_{l-1/2}}^{y_{l+1/2}} B_k(y) dy - h_{y1} \bar{B}_{kl} \bar{B}_{i+1,j} = \\ & h_{y1} \bar{B}_{kl} B_i(x_{j+1/2}). \end{aligned}$$

Daraus folgt die Stetigkeit am Interface $((x_{i+1/2}, y_{l-1/2}), (x_{i+1/2}, y_{l+1/2}))$.

Jetzt zur Stetigkeit der ersten Ableitung:

Da die Ableitung von D_{ik} nach x in y -Richtung konstant ist, können die Integrale bei den Stetigkeitsbetrachtungen weggelassen werden.

$$\left. \frac{\partial}{\partial x} D_{ik}(x_{j+1/2}, y_1) \right|_{\Omega_{j1}} = \bar{B}_{k1} B'_j(x_{i+1/2}) ,$$

$$\left. \frac{\partial}{\partial x} D_{ik}(x_{j+1/2}, y_1) \right|_{\Omega_{j+1,1}} = \bar{B}_{k1} B'_j(x_{i+1/2}) .$$

Völlig analog zeigt man die Interfaceeigenschaft in der x -Richtung.

□

2.1.14 Satz

Sei $f(x, y)|_{\partial\Omega} = 0$.

Dann ist
$$f(x, y) = \sum_{i,k=1}^n b_{ik} D_{ik}(x, y)$$

durch die Mittelpunktswerte U_{ik} eindeutig bestimmt.

Beweis:

Die Darstellung der Mittelpunktswerte mit Hilfe der D_{ik} ist ebenfalls durch das Tensorprodukt der eindimensionalen B-Splines möglich.

$$F_x = B_i(x_j)_{ij} = \frac{1}{8} \begin{bmatrix} 5 & 1 & & 0 \\ 1 & 6 & 1 & \\ & \cdot & \cdot & \cdot \\ 0 & & 1 & 6 & 1 \\ & & & 1 & 5 \end{bmatrix} \in \mathbb{R}^{n,n} , \quad F_y = F_x ,$$

$$H_x = \bar{B}_{k1} = \frac{1}{6} \begin{bmatrix} 3 & 1 & & 0 \\ 1 & 4 & 1 & \\ & \cdot & \cdot & \cdot \\ 0 & & 1 & 4 & 1 \\ & & & 1 & 3 \end{bmatrix} \in \mathbb{R}^{n,n} ,$$

Analoges gilt für F_y und H_y .

$$F := (D_{ik}(x_j, y_l))_{ikjl} = H_x \times F_x + H_y \times F_y - H_x \times H_y$$

$$= H_x \times \frac{1}{12} \begin{bmatrix} 9 & 1 & & 0 \\ 1 & 10 & 1 & \\ & \cdot & \cdot & \cdot \\ 0 & & 1 & 10 & 1 \\ & & & 1 & 9 \end{bmatrix} \quad \text{mit } H = H_x = H_y .$$

Nach Lemma 2.1.10 ist F positiv definit und damit folgt die Behauptung.

□

Damit ist auch einzusehen, daß D_{ik} eine Basis für $\mathbf{D}_n$ ist:

2.1.15 Satz

$D_{ik}(x,y)$ bilden eine Basis zu $\mathbf{D}_n$.

Beweis:

F aus Beweis 2.1.14 ist positiv definit $\Rightarrow \exists F^{-1}$

$$f(x_j, y_1) = \sum_{i,k=1}^n b_{ik} D_{ik}(x_j, y_1)$$

$\Rightarrow b_{ik} = F^{-1} f(x_j, y_1)$ damit ist jedes $u \in \mathbf{D}_n$ darstellbar. ☑

Zum Abschluß des Kapitels werden noch einige Matrixabschätzungen gezeigt, die für den Laplace-Operator von Wichtigkeit sein werden.

2.1.16 Satz

Sei $A_1 := \left(\int_{\Omega_{lm}} -k_1 \frac{\partial^2}{\partial x^2} B_{ik}(x,y) - k_2 \frac{\partial^2}{\partial y^2} B_{ik}(x,y) d(x,y) \right)_{iklm}$ bzw.

$\left(\int_{\Omega_{lm}} -k_1 \frac{\partial}{\partial x^2} D_{ik}(x,y) - k_2 \frac{\partial}{\partial y^2} D_{ik}(x,y) d(x,y) \right)_{iklm}$ mit

$B_{ik} \in \mathbf{S}_n$ bzw. $D_{ik} \in \mathbf{D}_n$.

Dann gilt für alle $\mathbf{b} \in \mathbb{R}^{n^2}$:

$$\mathbf{b}^T A_1 \mathbf{b} \geq c_1 h^2 (k_1 + k_2) \|\mathbf{b}\|_2^2.$$

Beweis:

Sei $A_x := \left(\int_{\Omega_l} -k_1 \frac{\partial^2}{\partial x^2} B_l(x) dx \right)_{il}$ wobei

$$A_x = -k_1 \frac{1}{h} \begin{bmatrix} 3 & -1 & & 0 \\ -1 & 2 & -1 & \\ & \cdot & \cdot & \cdot \\ 0 & -1 & 2 & -1 \\ & & -1 & 3 \end{bmatrix}$$

und $A_y := \left(\int_{\Omega_m} -k_2 \frac{\partial^2}{\partial y^2} B_k(y) dy \right)_{km}$ wobei

$$A_y = -k_2 \frac{1}{h} \begin{bmatrix} 3 & -1 & & 0 \\ -1 & 2 & -1 & \\ & \cdot & \cdot & \cdot \\ 0 & -1 & 2 & -1 \\ & & -1 & 3 \end{bmatrix}$$

Eine untere Schranke für die Eigenwerte von A_x bzw. A_y ist nach [STO-73] $\frac{3hk_1}{4}$ bzw. $\frac{3hk_2}{4}$. Das Tensorprodukt mit H liefert zusammen mit Lemma 2.1.10 die obige Behauptung. ☐

2.1.17 Satz

Sei $A_2 := \left(\int_{\Omega_{lm}} B_{ik}(x,y) d(x,y) \right)_{iklm}$ bzw.

$$A_2 := \left(\int_{\Omega_{lm}} D_{ik}(x,y) d(x,y) \right)_{iklm}.$$

Dann gilt für alle $\mathbf{b} \in \mathbb{R}^{n^2}$:

$$\mathbf{b}^T A_2 \mathbf{b} \geq c_2 h^2 \|\mathbf{b}\|_2^2.$$

Beweis:

$$\begin{aligned} A_2 &= h^2 (F_y \times H_x + F_x \times H_y - F_x \times F_y) \quad \text{bzw.} \\ &= h^2 (H_x \times H_y + H_y \times H_x - H_x \times H_y). \end{aligned}$$

Daraus folgt wieder mit Lemma 2.1.10 die Behauptung. ☐

2.1.18 Satz

Sei $A_3 := \left(\int_{\Omega} B_{ik}(x,y) B_{lm}(x,y) d(x,y) \right)_{iklm}$.

Dann gilt für alle $\mathbf{b} \in \mathbb{R}^{n^2}$:

$$c_{31} h^2 \|\mathbf{b}\|_2^2 \geq \mathbf{b}^T A_3 \mathbf{b} \geq c_{32} h^2 \|\mathbf{b}\|_2^2.$$

Beweis:

Zur Abschätzung werden drei verschiedene Typen von Matrizen benötigt:

1. Typ: $\left(\int_{\Omega_x} B_i(x_1) B_j(x_1) dx \right)_{ij} =: D_1$, wobei

$$D_1 = \frac{h^2}{960} \begin{bmatrix} 390 & 165 & 15 & 0 & 0 & 0 \\ 165 & 570 & 180 & 15 & 0 & 0 \\ 15 & 180 & 570 & 180 & 15 & 0 \\ 0 & 0 & 15 & 180 & 570 & 180 \\ 0 & 0 & 15 & 180 & 570 & 165 \\ 0 & 0 & 15 & 165 & 390 & 0 \end{bmatrix}$$

2. Typ : $(\int_{\Omega_x} B_i(x) B_j(x) dx)_{ij} =: D_2$, wobei

$$D_2 = \frac{h^2}{960} \begin{bmatrix} 320 & 180 & 20 & 0 & 0 & 0 & 0 & 0 \\ 180 & 520 & 200 & 20 & 0 & 0 & 0 & 0 \\ 20 & 200 & 520 & 200 & 20 & 0 & 0 & 0 \\ 0 & 0 & 20 & 200 & 520 & 200 & 20 & 0 \\ 0 & 0 & 0 & 20 & 200 & 520 & 200 & 180 \\ 0 & 0 & 0 & 20 & 180 & 320 & 0 & 0 \end{bmatrix}$$

3. Typ : $(\int_{\Omega_x} B_i(x) B_j(x) dx)_{ij} =: D_3$, wobei

$$D_3 = \frac{h^2}{960} \begin{bmatrix} 320 & 144 & 8 & 0 & 0 & 0 & 0 & 0 \\ 144 & 552 & 208 & 8 & 0 & 0 & 0 & 0 \\ 8 & 208 & 552 & 208 & 8 & 0 & 0 & 0 \\ 0 & 0 & 8 & 208 & 552 & 208 & 8 & 0 \\ 0 & 0 & 0 & 8 & 208 & 552 & 144 & 8 \\ 0 & 0 & 0 & 8 & 144 & 320 & 0 & 0 \end{bmatrix}$$

Es sei hier wiederholt, daß aus Gründen einer einfachen Darstellung das Gitter in x - und y -Richtung äquidistant gewählt wurde, d.h.

$$h_x = h_y = \text{konstant} = \frac{1}{n_x} = \frac{1}{n_y}$$

sowie

$$\Omega = [a,b] \times [a,b].$$

Damit ist A_3 darstellbar durch:

$$\begin{aligned} A_3 &= 2D_1 \times D_3 + 2D_1 \times D_2 - 2D_1 \times D_2 - 2D_1 \times D_2 + D_1 \times D_1 \\ &= 2D_1 \times D_3 - 2D_1 \times D_2 + D_1 \times D_1 \\ &= D_0 \times D_1 \end{aligned} \quad \text{mit} \quad D_0 = 2D_3 + D_1 - 2D_2.$$

D_0 ist positiv definit, mit Lemma 2.1.10 also auch A_3 . Daraus folgt die Behauptung. $\square$

Das Analogon zum vorherigen Satz für Ansatzfunktionen aus $\mathbf{D}_n$ ist:

2.1.19 Satz Sei $A_4 := (\int_{\Omega} D_{ik}(x,y) D_{lm}(x,y) d(x,y))_{iklm}$.

Dann gilt für alle $\mathbf{b} \in \mathbb{R}^{n^2}$:

$$h^2 \|\mathbf{b}\|_2^2 \leq \mathbf{b}^T A_4 \mathbf{b} \leq c_4 h^2 \|\mathbf{b}\|_2^2.$$

Beweis:

Zur Abschätzung wird ein weiterer Matrizen Typ gebraucht:

4. Typ : $\left(\int_{\Omega_x} B_1(x) B_j(x) dx \right)_{ij} =: D_4$, wobei

$$D_4 = \frac{h^2}{36} \begin{bmatrix} 9 & 3 & 1 & 0 & & & \\ 3 & 16 & 4 & 1 & 0 & & \\ 1 & 4 & 16 & 4 & 1 & 0 & \\ 0 & & & & & & 0 \\ & & 0 & 1 & 4 & 16 & 4 & 1 \\ & & & 0 & 1 & 4 & 16 & 3 \\ & & & & 0 & 1 & 3 & 9 \end{bmatrix}$$

A_3 läßt sich, falls man die gleichen Einschränkungen wie oben annimmt, sehr einfach darstellen:

$$A_3 = 5D_4 \times D_4 - 4D_4 \times D_4 = D_4 \times D_4.$$

Mit Lemma 2.1.10 folgt wiederum die Behauptung. ☐

Damit sind die Matrizenabschätzungen abgeschlossen und wir wenden uns dem zweidimensionalen Randwertproblem zu.

2.2 Das zweidimensionale Randwertproblem

2.2.1 Definition

Gegeben sei das zweidimensionale Randwertproblem:

$$\begin{aligned} \mathbf{L}u &= r & \text{auf } \Omega := [a,b] \times [a,b], \quad r \in L_2(\Omega), \\ & & \text{mit homogenen Randbedingungen} \end{aligned}$$

$$u|_{\partial\Omega} = 0 \quad \text{und}$$

$$\mathbf{L} : \mathbf{C}^2(\Omega) \longrightarrow \mathbf{C}(\Omega),$$

$$\mathbf{L}(x,y) := -k_1 u_{xx} - k_2 u_{yy}, \text{ wobei } k_1 \text{ und } k_2 \text{ positiv und konstant sind.}$$

Da die zweidimensionalen Ansatzfunktionen zu "nichtkonformen Elementen" führen, ist eine Darstellung des Laplace-Operators auf Räumen stückweise stetiger Funktionen nötig. Hierzu benötigt man eine Erweiterung des Laplace-Operators auf $\mathbf{S}_n$ bzw. $\mathbf{D}_n$. Dazu nun die folgende Definition:

2.2.2 Definition

Wie schon öfter erwähnt, soll aus Gründen der einfacheren Darstellung das Gebiet Ω in äquidistante Boxen unterteilt werden, sodaß die Kantenlänge der quadratischen Box $h = \frac{1}{n}$ ist, mit n als Anzahl der Unterteilungen in x - bzw. y -Richtung. Die Box Ω_{ik} beschreibt somit die i -te Box in x -Richtung und k -te in y -Richtung. Damit ist der Raum der j mal stückweise stetigen Funktionen auf Ω_{ik} definierbar:

$$\Omega_1 := (x_{1-1/2}, x_{1+1/2}), \quad \Omega_k := (y_{k-1/2}, y_{k+1/2}), \quad \Omega_{ik} := \Omega_1 \times \Omega_k,$$

$$\Omega_z := \bigcup_{i,k=1,\dots,n} \Omega_{ik},$$

$$\mathbf{C}_n^j(\Omega_z) := \{ u: \Omega_z \longrightarrow \mathbb{R} : u|_{\Omega_{ik}} \in \mathbf{C}^j(\Omega_{ik}) \quad \forall i,k \in \{1,\dots,n\} \}$$

Die zweidimensionale Partition hat die Form:

$$\blacksquare := \mathbf{t}_x \times \mathbf{t}_y \quad \text{wobei} \quad \mathbf{t}_x := \mathbf{t}_y := \mathbf{t}.$$

Auf $\Omega \setminus \blacksquare$ wird definiert:

$$\begin{aligned} L_x : \mathbf{C}_n^2(\Omega_z) &\longrightarrow \mathbf{C}_n(\Omega_z) & \text{mit} & \quad L_x u := -k_1 u_{xx} \\ L_y : \mathbf{C}_n^2(\Omega_z) &\longrightarrow \mathbf{C}_n(\Omega_z) & \text{mit} & \quad L_y u := -k_2 u_{yy} \\ L_{\blacksquare} : \mathbf{C}_n^2(\Omega_z) &\longrightarrow \mathbf{C}_n(\Omega_z) & \text{mit} & \quad L_{\blacksquare} := L_x + L_y. \end{aligned}$$

Für $u \in \mathbf{C}^2(\Omega)$ werden $L_x, L_y, L_{\blacksquare}$ entsprechend auf ganz Ω definiert.

Damit kann man völlig analog zum eindimensionalen Fall die ASWR-Lösung definieren, mit analogem Testraum:

2.2.3 Definition

$$K_{ik}(x,y) := \begin{cases} 1 & \text{für } (x,y) \in \Omega_{ik} \\ 0 & \text{sonst} \end{cases}$$

Der Testraum für die zweidimensionale ASWR-Lösung ist dann

$$\mathbf{K}_n := \text{span} \{ K_{ik}(x,y) \mid i,k = 1, \dots, n \}.$$

2.2.4 Definition

$u_A \in \mathbf{S}_n$ bzw. $\mathbf{D}_n$ ist ASWR-Lösung, falls für alle $i,k = 1,\dots,n$ gilt:

$$\int_{\Omega_{ik}} (L_{\blacksquare} u_A(x,y) - r(x,y)) \, d(x,y) = 0 \quad \text{d.h.}$$

$$((L_{\blacksquare} u_A - r), v)_{L_2(\Omega)} = 0 \quad \forall v \in \mathbf{K}_n \quad \text{und}$$

u_A erfüllt die Dirichlet-Randbedingungen von 2.2.1.

Die Existenz- und Eindeutigkeitsaussagen sind in erster Linie durch die Matrixabschätzungen des letzten Absatzes zu erhalten.

2.2.5 Satz

Das Randwertproblem hat zu beliebiger rechter Seite $r \in L_2(\Omega)$ genau eine ASWR-Lösung $u_A \in \mathbf{S}_n$ bzw. $\mathbf{D}_n$.

Beweis:

$$u_A = \sum_{i,k=1}^n \beta_{ik} B_{ik} \quad \text{bzw.} \quad u_A = \sum_{i,k=1}^n \beta_{ik} D_{ik}.$$

Aufgrund der Definition der ASWR-Lösung läßt sich dies wie folgt darstellen:

$$(1) \quad A_1 \mathbf{B} = \mathbf{R} \quad \text{mit} \quad \mathbf{R} := \int_{\Omega_{ik}} r(x,y) d(x,y).$$

Für A_1 gilt nach Satz 2.1.16: $\mathbf{B}^T A_1 \mathbf{B} \geq c_1 h^2 \|\mathbf{B}\|_2^2 \Rightarrow \exists A_1^{-1}$

Damit folgt die Existenz und Eindeutigkeit von (1) und damit die Behauptung. ☐

Für die Konvergenzaussage soll zur besseren Übersicht die Beweisidee für die ASWR-Lösung aus $\mathbf{S}_n$ und $\mathbf{D}_n$ skizziert werden:

Zunächst wird ein $u_s \in \mathbf{S}_n$ durch die Superposition der jeweiligen eindimensionalen ASWR-Lösungen konstruiert. Für dieses u_s wird folgende Abschätzung angegeben:

$$\|u_s - u_0\|_{L_2(\Omega)} = O(h^2).$$

Danach wird die Konvergenz von u_s gegen die zweidimensionale ASWR-Lösung aus $\mathbf{S}_n$ nachgewiesen:

$$\|u_s - u_A\|_{L_2(\Omega)} = O(h^2).$$

Damit ist die Konvergenzaussage für die ASWR-Lösung aus $\mathbf{S}_n$ gesichert. Für die interessantere ASWR-Lösung aus $\mathbf{D}_n$ wird ein Operator definiert, der $u_s \in \mathbf{S}_n$ auf ein $u_{ss} \in \mathbf{D}_n$ abbildet. Diese Abbildung wird abgeschätzt, sodaß folgendes Ergebnis erhalten wird:

$$\|u_s - u_{ss}\|_{L_2(\Omega)} = O(h^2).$$

In einer weiteren Abschätzung

$$\|u_{ss} - u_{A2}\|_{L_2(\Omega)} = O(h^2)$$

mit $u_{A2} \in \mathbf{D}_n$ ist das eigentliche Ziel, die Konvergenz der ASWR-Lösung aus $\mathbf{D}_n$, gezeigt.

Für die Konstruktion eines $u_s \in \mathbf{S}_n$ durch Superposition der eindimensionalen Lösungen ist die Erweiterung der eindimensionalen Lösungen auf das zweidimensionale Gebiet nötig. Dazu die nächste Definition, die dies zum Inhalt hat.

2.2.6 Definition

Seien G_x, G_y die jeweiligen Greenschen Operatoren für die eindimensionalen Lösungen in x - bzw. y -Richtung und P_x, P_y die eindimensionalen Projektoren in den Raum der stückweise konstanten Funktionen in x - bzw. y -Richtung. Weiter seien L_x und L_y die jeweiligen für das eindimensionale Problem definierten Laplace-Operatoren. Dann ist die Erweiterung der eindimensionalen Lösungen auf den den zweidimensionalen Raum folgendermaßen definiert:

$$Xu(.,y) := G_x P_x \frac{1}{k_1} u_{xx}(.,y) \quad \forall x,y \in \Omega,$$

$$Yu(x,.) := G_y P_y \frac{1}{k_2} u_{yy}(x,.) \quad \forall x,y \in \Omega.$$

$Xu(.,y)$ bzw. $Yu(x,.)$ ist also ASWR-Lösung des eindimensionalen Problems in x bzw. y -Richtung.

Rein formal läßt sich die Abschätzung des eindimensionalen ASWR-Problems somit auch auf X und Y anwenden.

2.2.7 Abschätzungen

Für $u \in \mathbf{C}^{3,0}(\Omega)$ gilt:

$$\|(Xu - u)(.,y)\|_{L_\infty(\Omega)} \leq c_x h^2 \|D^{3,0}(.,y)\|_{L_\infty(\Omega)}.$$

Für $u \in \mathbf{C}^{0,3}(\Omega)$ gilt:

$$\|(Yu - u)(x,.)\|_{L_\infty(\Omega)} \leq c_y h^2 \|D^{0,3}(x,.)\|_{L_\infty(\Omega)}.$$

Im weiteren wird der Raum, der aus dem Produkt der quadratischen eindimensionalen B-Splines gebildet wird, vorgestellt. Danach wird eine Abbildung definiert, die aus dem oben angesprochenen Raum in den $\mathbf{S}_n$ führt. Eine Abschätzung dieser Abbildung zusammen mit dem Ergebnis der eindimensionalen ASWR-Abschätzung führt auf die Aussage $\|u_s - u_0\|_{L_2(\Omega)} = O(h^2)$. Zunächst also der Raum der eindimensionalen Produktsplines:

2.2.8 Definition

$$B_{ik}^*(x,y) := B_i(x) B_k(y) \quad \text{für } B_i(x) \in \mathbf{S}_{2,t}(\Omega_x) \quad \text{und} \\ B_k(y) \in \mathbf{S}_{2,t}(\Omega_y).$$

$$\bar{\mathbf{B}}_n := \text{span}(B_{ik}^*(x,y) \mid i,k = 1,..,n)$$

Die nächste Definition beschreibt die Abbildungsvorschrift von $\bar{\mathbf{B}}_n$ nach $\mathbf{S}_n$.

2.2.9 Definition

$W : \bar{B}_n(\Omega) \longrightarrow S_n(\Omega)$, mit folgender Abbildungsvorschrift

$$Wu := \sum_{j=0}^2 \frac{(x - x_1)^j}{j!} \frac{\partial^j}{\partial x^j} u(x_1, y_k) + \sum_{l=1}^2 \frac{(y - y_k)^l}{l!} \frac{\partial^l}{\partial y^l} u(x_1, y_k).$$

Wu ist die Taylor-Entwicklung um den Punkt (x_1, y_k) . Damit ist auch der nächste Satz einsichtig.

2.2.10 Satz

Für $u \in \bar{B}_n$ gilt:

$$\|Wu - u\|_{L_\infty(\Omega)} \leq c_W \sum_{i,k=1}^2 h^{i+k} \|D^{i,k}u\|_{L_\infty(\Omega)}.$$

Beweis:

Durch Taylorentwicklung um den Punkt (x_1, y_k) . ☐

Im letzten Satz wurde $u \in \bar{B}_n$ vorausgesetzt. Der nächste Satz konstruiert ein derartiges Element mit Hilfe eindimensionaler ASWR-Lösungen.

2.2.11 Satz

Für $u \in C^{2,2}(\Omega)$ ist $XYu = YXu \in \bar{B}_n$.

Beweis:

$$Xu = \sum_{i=1}^n b_i(y) B_i(x)$$

$$b_i(y) \in C^2(\Omega_y) \quad \text{da } Xu \in C^{1,2}(\Omega).$$

$$Yb_i(y) = \sum_{k=1}^n b_{i,k} B_k(y) \Rightarrow$$

$$XYu = \sum_{k=1}^n \sum_{i=1}^n b_{ik} B_i(x) B_k(y).$$
☐

Nun wird ein geeigneter Raum angegeben, aus der die wahre Lösung stammen soll. Dieser ist so konstruiert, daß gerade die jeweiligen eindimensionalen Lösungen darin enthalten sind.

2.2.12 Definition

$$\mathbf{C}^*(\Omega) := \mathbf{C}^{2,3}(\Omega) \cap \mathbf{C}^{3,2}(\Omega)$$

$$\|u\|_* := \max_{i,k=0,\dots,3, i+k \leq 5} \|D^{i,k} u\|_{L^\infty(\Omega)}.$$

Damit ist u_s aus $u \in \mathbf{C}^*(\Omega)$ durch die eindimensionalen Lösungen und den Operator W konstruierbar.

2.2.13 Satz

$$\begin{aligned} WXY : \mathbf{C}^*(\Omega) &\longrightarrow \mathbf{S}_n(\Omega) \text{ mit} \\ \|WXYu - u\|_{L^\infty(\Omega)} &= O(h^2) \end{aligned}$$

Beweis:

Dies folgt durch die Dreiecksungleichung aus den eindimensionalen Abschätzungen und Satz 2.2.10. ☑

Damit kommt man zur Konvergenz der ASWR-Lösung mit der ersten Interfacebedingung. Wie schon oben erwähnt, wird die Konvergenz der ASWR-Lösung über die Konvergenz der Superposition der eindimensionalen Lösungen, die ein u_s aus $\mathbf{S}_n$ konstruieren und der zweidimensionalen ASWR-Lösung aus $\mathbf{S}_n$ abgeleitet. Dies geschieht über ein Hilfsrandwertproblem, dessen rechte Seite $L_\square(u_s - u_A)$ ist und dessen Lösung somit $(u_s - u_A)$ ist. Folgender Satz, der die Lösung durch die rechte Seite abschätzt, wird zum Beweis für die Konvergenz von $\|u_s - u_A\| \longrightarrow 0$ benötigt.

2.2.14 Satz

Sei u_A ASWR-Lösung zum gegebenen Randwertproblem. Dann gilt:

$$\begin{aligned} \|u_A\|_{L_2(\Omega)} &\leq c_{rs} \|r\|_{L_2(\Omega)} \\ &\leq c_{rs} h^{-1} \sqrt{\sum_{i,k=1}^n \left(\int_{\Omega_{ik}} r(x,y) d(x,y) \right)^2}. \end{aligned}$$

Beweis:

$$\begin{aligned} \|R\|_2^2 &:= \sum_{i,k=1}^n \left(\int_{\Omega_{ik}} r(x,y) d(x,y) \right)^2 \\ &\leq \sum_{i,k=1}^n \int_{\Omega_{ik}} 1^2 d(x,y) \int_{\Omega_{ik}} (r(x,y))^2 d(x,y) \end{aligned}$$

$$(1) \quad \leq \sum_{i,k=1}^n h^2 \int_{\Omega_{ik}} (r(x,y))^2 d(x,y) = h^2 \|r\|_{L_2(\Omega)}^2$$

Für A_1 (Kopplung erster Art) gilt mit 2.1.16

$$\mathbf{B} = A_1^{-1} \mathbf{R}.$$

Wegen der Invertierbarkeit der Matrix A_1 ist

$$\begin{aligned} \|\mathbf{B}\|_2 &\leq \|A_1^{-1}\|_2 \|\mathbf{R}\|_2 \\ &\leq \frac{1}{c_1(k_1+k_2)} h^{-2} \|\mathbf{R}\|_2 \\ &\leq \frac{1}{c_1(k_1+k_2)} h^{-1} \|\mathbf{r}\|_2. \end{aligned}$$

Mit Satz 2.1.18 läßt sich die ASWR-Lösung durch $\|\mathbf{B}\|_2$ abschätzen und somit gilt die Behauptung. □

Jetzt kann man mit dem Hilfsrandwertproblem die Konvergenz von u_s an u_A zeigen.

2.2.15 Satz

Sei u_s die Lösung durch Superposition der eindimensionalen ASWR-Lösungen, also $u_s = WXYu_0$ aus $\mathbf{S}_n$ mit u_0 Lösung aus $\mathbf{C}^*$. Dann gilt:

$$\|u_s - u_A\|_{L_2(\Omega)} \leq c_* h^2 \|u_0\|_*.$$

Beweis:

Formulierung des Hilfsrandwertproblems:

$$\begin{aligned} L_{\blacksquare}(u) &= r_h \quad \text{wobei } r_h := L_{\blacksquare}(u_s - u_A) \in L_2(\Omega), \\ u|_{\partial\Omega} &= 0. \end{aligned}$$

Es ist $(u_s - u_A)$ ASWR-Lösung zum Hilfsrandwertproblem, da $(u_s - u_A) \in \mathbf{S}_n$ ist. Nach dem letzten Satz 2.2.14 läßt sich die Lösung durch die rechte Seite abschätzen:

$$(1) \quad \|u_s - u_A\|_{L_2(\Omega)} \leq c_{rs} h^{-1} \|(r_h)_{i,k=1,\dots,n}\|_2 \quad \text{wobei}$$

$$(r_h)_{i,k} := \int_{\Omega_{ik}} L_{\blacksquare}(u_s - u_A)(x,y) \, d(x,y)$$

$$\text{Nun ist } |(r_h)_{i,k}| = \left| \int_{\Omega_{ik}} L_{\blacksquare}(WXYu_0 - u_A)(x,y) \, d(x,y) \right|$$

$$(2) \quad = \left| \int_{\Omega_{ik}} L_{\blacksquare}(WXYu_0 - u_0)(x,y) \, d(x,y) \right| \quad (\text{wegen ASWR-Definition}).$$

Weiter ist

$$\begin{aligned}
 \int_{\Omega_{ik}} L_x(XY u_0 - u_0)(x, y) \, d(x, y) &= \\
 \int_{\Omega_k} \int_{\Omega_i} L_x(XY u_0 - u_0)(x, y) \, dx dy &= \\
 \int_{\Omega_k} \int_{\Omega_i} (L_x XY u_0 - L_x u_0)(x, y) \, dx dy &= \\
 \int_{\Omega_k} \int_{\Omega_i} (L_x Y u_0 - L_x u_0)(x, y) \, dx dy &\leq \quad (\text{wegen ASWR-Definition}) \\
 h^2 \|L_x Y u_0 - L_x u_0\|_{L_\infty(\Omega)} &\leq \\
 c_x h^4 \|D^{0,3} L_x u_0\|_{L_\infty(\Omega)} &\leq \\
 h^4 (c_x \|u_0\|_* + c_y \|u_0\|_*) \quad , \quad \text{da} \\
 \|D^{0,3} L_x u_0\|_{L_\infty(\Omega)} &\leq c_d \max_{j=0,1,2} \|D^{j,3} u_0\|_{L_\infty(\Omega)} \quad .
 \end{aligned}$$

Analoges erhält man für $\int_{\Omega_{ik}} L_y(XY u_0 - u_0)(x, y) \, d(x, y)$, sodaß

$$(3) \quad \max_{i,k=1,n} \left| \int_{\Omega_{ik}} L_y(XY u_0 - u_0)(x, y) \, d(x, y) \right| \leq c_z h^4 \|u_0\|_* \quad \text{ist.}$$

Nun wird eine analoge Abschätzung für den Operator W definiert:

$$\begin{aligned}
 (u - Wu)(x, y) &= \sum_{l,m=1}^2 (x - x_l)^l (y - y_m)^m c_{lm} \quad \text{mit } u \in \mathbf{B}_n \text{ und} \\
 c_{lm} &:= \frac{1}{l \cdot m} D^{l,m} u(x_l, y_m) \quad , \quad (x, y) \in \Omega_{ik} \quad .
 \end{aligned}$$

Weiter ist

$$\begin{aligned}
 (4) \quad &\left| \int_{x_{l-1/2}}^{x_{l+1/2}} \int_{y_{m-1/2}}^{y_{m+1/2}} L_x(Wu - u)(x, y) \, d(x, y) \right| = \\
 &\left| \int_{x_{l-1/2}}^{x_{l+1/2}} \int_{y_{m-1/2}}^{y_{m+1/2}} k_1 2 \sum_{m=1}^2 (y - y_m)^m c_{2m} \, d(x, y) \right| = \\
 &\left| \frac{1}{6} h^3 \int_{x_{l-1/2}}^{x_{l+1/2}} k_1 c_{2m} \, d(x, y) \right| \leq \\
 &c_w h^4 \|u\|_* \quad .
 \end{aligned}$$

(3) und (4) ergeben mit Hilfe der Dreiecksungleichung angewandt auf (2) folgendes:

$$(5) \quad |(r_h)_{i,k}| \leq c_h h^4 \|u_0\|_*$$

Nun ist $\|(r_h)_{i,k=1,\dots,n}\|_2 \leq h^{-1} |(r_h)_{i,k}|$, so daß

$$\begin{aligned}
 \|u_S - u_A\|_{L_2(\Omega)} &\leq c_{rs} h^{-2} |(r_h)_{i,k}| \quad \text{mit (5)} \\
 &\leq c_{rs} h^{-2} c_h h^4 \|u_0\|_* = O(h^2) \quad \text{ist.}
 \end{aligned}$$

□

Damit ist mit Hilfe der Dreiecksungleichung aus Satz 2.2.13 und 2.2.15 die Konvergenz für die zweidimensionale ASWR-Lösung aus $\mathbf{S}_n$ bewiesen. Der nächste Abschnitt hat die ASWR-Lösung aus $\mathbf{D}_n$ zum Inhalt.

2.3 Konvergenz des ASWR-Verfahrens mit neuen Interface-Bedingungen

Die Idee des Konvergenzbeweises liegt darin, einen Operator zu finden, der eine ASWR-Lösung aus $\mathbf{S}_n$ auf ein Element aus $\mathbf{D}_n$ abbildet. Daraufhin zeigt man die Konvergenz dieses Elementes gegen die ASWR-Lösung aus $\mathbf{D}_n$. Da alle anderen Instrumentarien für den Beweis wie z.B. eine B-Splinebasis im letzten Kapitel schon ausgeführt worden sind, bleibt hier nur noch eine Abschätzung für diesen Abbildungsoperator zu zeigen.

2.3.1. Definition

Sei $T : \mathbf{S}_n \longrightarrow \mathbf{D}_n$, mit
 $s \longmapsto d$, wobei folgende Abbildungsvorschrift gilt:
 $s(x_i, y_k) = T(s(x_i, y_k)) = d(x_i, y_k)$ für alle $i, k=1, \dots, n$ und
 $T(s)$ genüge den gleichen Randbedingungen wie s .

2.3.2 Bemerkung

Die Abbildung T ist bijektiv, da sowohl s wie auch d durch die Mittelpunkts-
 werte (x_i, y_k) und die Randwerte (Satz 2.1.14) eindeutig bestimmt sind.

Diesen Operator T gilt es nun abzuschätzen, um die Konvergenz zwischen der ASWR-Lösung aus $\mathbf{S}_n$ und d , wie oben definiert, zu zeigen.

2.3.3 Satz

Für $u \in \mathbf{S}_n$ gilt:

$$\|Tu - u\|_{L_\infty(\Omega \setminus \Gamma_k)} = O(h_1^2) + O(h_k^2)$$

Beweis:

Man betrachtet das jeweilige Polynom auf der Box i_k :

$$P_{i,k}(x,y) = U_{i,k} +$$

$$\begin{aligned} & \frac{(x - x_1)}{h_1} (U_{i+1/2,k} - U_{i-1/2,k}) + \frac{(x - x_1)^2}{h_1^2} 2(U_{i+1/2,k} + U_{i-1/2,k} - 2U_{i,k}) \\ & + \frac{(y - y_k)}{h_k} (U_{i,k+1/2} - U_{i,k-1/2}) + \frac{(x - x_k)^2}{h_k^2} 2(U_{i,k-1/2} + U_{i,k+1/2} - 2U_{i,k}) \end{aligned}$$

aus $\mathbf{S}_n$ und

$$P_{i,k}^O(x,y) = V_{i,k} +$$

$$\begin{aligned} & \frac{(x - x_1)}{h_1} (V_{i+1/2,k} - V_{i-1/2,k}) + \frac{(x - x_1)^2}{h_1^2} 2(V_{i+1/2,k} + V_{i-1/2,k} - 2V_{i,k}) \\ & + \frac{(y - y_k)}{h_k} (V_{i,k+1/2} - V_{i,k-1/2}) + \frac{(x - x_k)^2}{h_k^2} 2(V_{i,k-1/2} + V_{i,k+1/2} - 2V_{i,k}) \end{aligned}$$

aus $\mathbf{D}_n$.

Nun sei $T(P_{i,k}(x,y)) = P_{i,k}^O(x,y)$. Dann gilt $U_{i,k} = V_{i,k}$.

Außerdem gilt: $P_{i,k}(x_{i+1/2}, y_k) = P_{i+1,k}(x_{i+1/2}, y_k)$,

$$\int_{y_{k-1/2}}^{y_{k+1/2}} P_{i,k}^O(x_{i+1/2}, y) dy = \int_{y_{k-1/2}}^{y_{k+1/2}} P_{i+1,k}^O(x_{i+1/2}, y) dy \quad (\Rightarrow)$$

$$V_{i+1/2,k} + \frac{1}{6} (V_{i,k-1/2} + V_{i,k+1/2} - 2V_{i,k}) =$$

$$V_{i+1-1/2,k} + \frac{1}{6} (V_{i+1,k-1/2} + V_{i+1,k+1/2} - 2V_{i+1,k}).$$

Die Boxennachbarnpunkte $V_{i+1/2,k}$ und $V_{i+1-1/2,k}$ sind am Interface ungleich; die Unstetigkeit wird also durch die verschiedenen Werte in den Klammern bestimmt. Für die Klammern aber gilt:

$$\frac{1}{6} (V_{i,k-1/2} + V_{i,k+1/2} - 2V_{i,k}) = \frac{h_k^2}{24} \frac{\partial^2}{\partial y^2} P_{i,k}^O \quad \text{analog}$$

$$\frac{1}{6} (V_{i+1,k-1/2} + V_{i+1,k+1/2} - 2V_{i+1,k}) = \frac{h_k^2}{24} \frac{\partial^2}{\partial y^2} P_{i+1,k}^O$$

$$\Rightarrow |V_{i+1/2,k} - V_{i+1-1/2,k}| = O(h_k^2).$$

Falls aber $V_{i+1/2,k} = V_{i+1-1/2,k}$ so gilt $V_{i+1/2,k} = V_{i+1-1/2,k} = U_{i+1/2,k}$.

Berücksichtigt man, daß $P_{i,k}(x,y)$ und $P_{i,k}(x,y)$ am Rand identisch sind, so gilt:

$$(I) \quad |V_{i+1/2,k} - U_{i+1/2,k}| = O(h_k^2)$$

Dann erhält man:

$$\begin{aligned} & \|P_{i,k}(x,y) - P_{i,k}^O(x,y)\|_{L_\infty(\Omega)_k} = \\ & \left\| \frac{(x - x_i)}{h_i} ((U_{i+1/2,k} - U_{i-1/2,k}) - (V_{i+1/2,k} - V_{i-1/2,k})) \right. \\ & \quad \frac{(x - x_i)^2}{h_i^2} 2((U_{i+1/2,k} + U_{i-1/2,k} - 2U_{i,k}) - (V_{i+1/2,k} + V_{i-1/2,k} - 2V_{i,k})) \\ & \quad + \frac{(y - y_k)}{h_k} ((U_{i,k+1/2} - U_{i,k-1/2}) - (V_{i,k+1/2} - V_{i,k-1/2})) + \\ & \quad \left. \frac{(x - x_k)^2}{h_k^2} 2((U_{i,k-1/2} + U_{i,k+1/2} - 2U_{i,k}) - (V_{i,k-1/2} + V_{i,k+1/2} - 2V_{i,k})) \right\| \\ & = O(h_i^2) + O(h_k^2) \quad \text{wegen (I).} \end{aligned}$$



2.3.4 Lemma

Es gilt weiter: $\|L_X(P_{i,k}(x,y) - P_{i,k}^O(x,y))\|_{L_\infty(\Omega)_k} = O(h_k^2)$.

Beweis:

$$\begin{aligned} & \|L_X(P_{i,k}(x,y) - P_{i,k}^O(x,y))\|_{L_\infty(\Omega)_k} = \\ & \left(\frac{4}{h_i^2} ((U_{i+1/2,k} + U_{i-1/2,k} - 2U_{i,k}) - (V_{i+1/2,k} + V_{i-1/2,k} - 2V_{i,k})) \right) \\ & = O(h_k^2) \quad (\text{wegen 2.3.3 (I)}). \end{aligned}$$



Ähnlich wie in Satz 2.2.15 wird nun zur Konvergenzuntersuchung ein Hilfsrandwertproblem betrachtet:

2.3.5 Satz

Sei $u_{ss} := TWXYu_0 \in \mathbf{D}_n$ die u_0 der Lösung aus $\mathbf{C}^*(\Omega)$ und u_{A_2} die ASWR-Lösung aus $\mathbf{D}_n$. Dann gilt:

$$\|u_{ss} - u_{A_2}\|_{L_2(\Omega)} = O(h^2).$$

Beweis:

Formulierung des Hilfsrandwertproblems:

$$\begin{aligned} L_\square(u) &= r_h \quad \text{wobei } r_h := L_\square(u_{ss} - u_{A_2}) \in L_2(\Omega), \\ u|_{\partial\Omega} &= 0. \end{aligned}$$

Wegen $(u_{ss} - u_{A2}) \in \mathbf{D}_n$ ist $(u_{ss} - u_{A2})$ ASWR-Lösung zum Hilfsrandwertproblem. Nach dem Satz 2.2.14 läßt sich die Lösung durch die rechte Seite abschätzen:

$$\|u_{ss} - u_{A2}\|_{L_2(\Omega)} \leq c_{rs} h^{-1} \|(r_h)_{i,k=1,\dots,n}\|_2 \quad \text{wobei}$$

$$(1) \quad (r_h)_{i,k} := \int_{\Omega_{ik}} L_{\mathbf{m}}(u_{ss} - u_{A2})(x,y) \, d(x,y)$$

Nun ist

$$(2) \quad \int_{x_{l-1/2}}^{x_{l+1/2}} \int_{y_{m-1/2}}^{y_{m+1/2}} L_x(Tu - u)(x,y) \, d(x,y) \leq h_l h_m \|L_x Tu - L_x u\|_{L_\infty(\Omega_{ik})} = O(h^4), \quad \text{für } h_l = h_m = h.$$

(wegen Lemma 2.3.4).

Für (1) gilt nun wegen (2), 2.2.15 (3), 2.2.15(4) und der Dreiecksungleichung:

$$(3) \quad |(r_h)_{i,k}| = O(h^4)$$

Nun ist $\|(r_h)_{i,k=1,\dots,n}\|_2 \leq h^{-1} |(r_h)_{i,k}|$, so daß

$$\begin{aligned} \|u_{ss} - u_A\|_{L_2(\Omega)} &\leq c_{rs} h^{-2} |(r_h)_{i,k}| \quad \text{mit (3)} \\ &= O(h^2) \quad \text{ist.} \end{aligned}$$



Damit kommt man nun zu dem abschließenden Satz des Kapitels, der ein Hauptergebnis der Arbeit darstellt:

2.3.6 Satz

Sei u_0 die Lösung aus $\mathbf{C}^*(\Omega)$ zum gegebenen Randwertproblem. Dann gilt für die ASWR-Lösungen u_A aus $\mathbf{S}_n(\Omega)$ und $\mathbf{D}_n(\Omega)$:

$$\|u_A - u_0\|_{L_2(\Omega)} \leq c_A h^2 \|u_0\|_*.$$

Beweis:

Mit Hilfe der Dreiecksungleichung und den Ergebnissen aus den Sätzen 2.2.13, 2.2.15, 2.3.3 und 2.3.5 sind die Behauptungen gezeigt.



3 Zeitabhängiges Problem

Simulation der Festkörperdiffusion führt auf Systeme parabolischer Differentialgleichungen, vgl. Kapitel 4 für die physikalische Modellierung.

In diesem Kapitel wird die Konvergenz von ASWR-Lösungen für eine lineare parabolische Differentialgleichung bewiesen, wobei Methoden aus [ODE-76], [BAK-76], [SCH-83], [SVO-83] herangezogen werden.

3.1 Parabolisches Problem

Das hier betrachtete Problem ist von parabolischem Typ. Es unterscheidet sich vom elliptischen Problem 2.2.1 nur durch Hinzunahme einer Zeitableitung. Für die Behandlung des Zeitverlaufs ist folgende Definition nötig:

3.1.1 Definition

Sei X ein linearer, normierter Raum, I ein nichtentartetes Intervall aus $\mathbb{R}$.

Eine Abbildung $f : I \rightarrow X$ heißt differenzierbar mit Ableitung $a = \frac{df}{dt}$ im Punkt $t_p \in I$ ($\Rightarrow$)

$$\exists a \in X \quad \forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall t \in K_\delta(t_p) \cap I \setminus \{t_p\} : \quad \left\| a - \frac{f(t) - f(t_p)}{t - t_p} \right\|_X < \varepsilon.$$

$$C(I, X) := \{f \mid f \text{ ist stetig}\}$$

$$C^k(I, X) := \{f \in C^{k-1}(I, X) \mid \exists f^{(k)} \text{ und } f^{(k)} \in C(I, X)\}$$

Nach [LAD-68] werden parabolische Systeme folgendermaßen klassifiziert:

3.1.2 Definition

Das Differentialgleichungssystem mit m Gleichungen

$$u_t - \frac{\partial}{\partial x_i} (a_{ij}(x, t) u_{x_j}) = r, \quad u \in \mathbb{R}^m, \quad i, j = 1, \dots, m$$

ist von parabolischem Typ, falls für beliebiges $\zeta \in \mathbb{R}^m$ gilt:

$$\nu \zeta^2 \leq a_{ij} \zeta_i \zeta_j \leq \omega \zeta^2, \quad \nu > 0.$$

3.1.3 Definition

Sei $u(t) := u(x, y, t)$ für $(x, y) \in \Omega$, $t \in [t_0, t_e]$

$$\frac{d}{dt}u(t) + \mathbf{L}u(t) = r(t), \quad t_0 < t < t_e \quad \text{mit } r(t) \in C^1([t_0, t_e], \mathbf{C}(\Omega)).$$

Die Anfangsbedingung sei

$$u_0 := u(t_0) := g \in C^1([t_0, t_e], \mathbf{C}^2(\Omega)).$$

$\mathbf{L}$ sei wie in 2.2.1 definiert mit analogen Randbedingungen.

Wir setzen eine für die Konvergenzaussage notwendige Bedingung voraus: es existiere eine Lösung aus $C^1([t_0, t_e], \mathbf{C}^*(\Omega))$. Weiterhin benötigen wir eine der Koerzitivität des Laplace-Operators $\mathbf{L}_\square$ entsprechende Bedingung.

3.1.4 Definition

$\mathbf{L}_\square$ sei definiert wie 2.2.2.

$$\begin{aligned} \text{Sei } B_\square : \mathbf{C}_n^2(\Omega \setminus \partial \square) \times \mathbf{C}_n(\Omega \setminus \partial \square) &\longrightarrow \mathbb{R} \\ (u, v) &\longmapsto \int_{\Omega} \mathbf{L}_\square u(x, y) v(x, y) \, d(x, y) \end{aligned}$$

$\mathbf{L}_\square$ ist α -koerzitiv, falls $B_\square(u, u) \geq c_{\text{const}} \|u'\|_{L_2(\Omega)}^2$

Für den Operator $\mathbf{L}$ aus 2.1.1 folgt die α -Koerzitivität mit Hilfe der Lagrange-Identität aus der Koerzitivität.

Analog zum Projektor $\mathbf{P}$ in 1.4.8 läßt sich ein Operator $\mathbf{P}$ definieren, der auf $\mathbf{K}_n$ projiziert.

3.1.5 Definition

Sei $\mathbf{P}$ der lineare Projektor in den Raum der stückweise konstanten Funktionen $\mathbf{K}_n$

$$\begin{aligned} \mathbf{P} : L_2(\Omega) &\longrightarrow \mathbf{K}_n, \\ f &\longmapsto \mathbf{P}f, \quad \text{mit} \\ (1) \quad \int_{\Omega} (f - \mathbf{P}f) g \, d(x, y) &= 0 \quad \forall g \in \mathbf{K}_n \end{aligned}$$

$\mathbf{P}f$ wird die L_2 -Approximation von f aus $\mathbf{K}_n$ genannt.

Die a -Koerzivität ist dem asymmetrischen Galerkin-Verfahren angepaßt. Eine weitere Anpassung der Koerzivität an das ASWR-Verfahren bringt die folgende Definition:

3.1.6 Definition

$L_{\mathbf{n}}$ sei definiert wie 2.2.2

$L_{\mathbf{n}}$ ist ASWR-koerzitiv, falls $B_{\mathbf{n}}(u, Pu) \geq c_{rr} \|u\|_{L_2(\Omega)}^2 \quad \forall u \in \mathbf{D}_{\mathbf{n}}$.

Die ASWR-Koerzivität kann aus der a -Koerzivität mit Hilfe der Matrixabschätzungen aus 2.1.14 - 2.1.17 und der Ungleichung von Rayleigh-Ritz-Wirtunger geschlossen werden.

3.2 Existenz der ASWR-Lösung

Die Definition der ASWR-Lösung läßt sich analog zu 2.2.4 festlegen.

3.2.1 Definition

Sei $g_A \in \mathbf{D}_{\mathbf{n}}$ die "Startnäherung" an g . Dann heißt $u_A \in C^1([t_0, t_e], \mathbf{D}_{\mathbf{n}})$ semidiskrete ASWR-Lösung des Anfangsrandwertproblems 3.1.3, falls gilt:

$$u_A(0) = g_A$$

$$u_A(t) \text{ erfüllt die Randbedingungen für } t \in [t_0, t_e]$$

$$\left(\frac{d}{dt} u_A(t) + \mathbf{L} u_A(t) - r(t) \right), v, \quad \forall v \in \mathbf{K}_{\mathbf{n}}, \quad t_0 < t < t_e.$$

Analog zu Satz 2.2.5 zeigt man die Existenz der ASWR-Lösung

3.2.2 Satz

Zu einem gegebenen $g_A \in \mathbf{D}_{\mathbf{n}}$ existiert genau eine semidiskrete ASWR-Lösung:

Beweis:

Sei $R_{ik}(t) := (r(t), K_{ik})_{L_2(\Omega)}$, $K_{ik} \in \mathbf{K}_{\mathbf{n}}$

$$R(t) := (R_{ik}(t) \mid i, k = 1, \dots, n).$$

$$R \in C^1([t_0, t_e], \mathbb{R}^{n^2}) \text{ da } r \in C^1([t_0, t_e], \mathbf{C}(\Omega)).$$

Die semidiskrete ASWR-Lösung läßt sich durch eine Linearkombination der

Basiselemente darstellen: $u_A(t) = \sum_{i,k=1}^n \beta(t)_{ik} D_{ik}$.

$\beta \in C^1([t_0, t_e], \mathbb{R}^{n^2})$ und lassen sich aus folgender Matrixgleichung bestimmen: $A_2 \frac{d}{dt} \beta(t) + A_1 \beta(t) - R(t) = 0$. Aus 2.1.16 und 2.1.17 folgt die Invertierbarkeit von A_1 und A_2 und damit die Behauptung.



3.3 Konvergenz der ASWR-Lösung

In der nächsten Definition wird ein Operator Q konstruiert, der die Lösung u von $\mathbf{L}u - r = 0$ auf die zugehörige ASWR-Lösung aus $\mathbf{D}_n$ projiziert. Die rechte Seite des Laplace-Problems wird gerade so definiert, daß $Qu = r$ ist. Dies ist möglich, da es zu $r \in L_2(\Omega)$ eine ASWR-Lösung gibt.

3.3.1 Definition

Sei $Q : W_{2,\infty}(\Omega \setminus \bar{\Omega}) \longrightarrow \mathbf{D}_n(\Omega \setminus \bar{\Omega})$

$u \longmapsto Qu$ definiert durch:

Qu ist ASWR-Lösung vom elliptischen Problem $\mathbf{L}u = r$, wobei Qu die Randbedingungen erfülle.

Aus der Matrixdarstellung von 2.1.16 ist Linearität und Stetigkeit des Operators ersichtlich. Somit ist $Qu \in C^1([t_0, t_e], \mathbf{D}_n)$, falls $u \in C^1([t_0, t_e], W_{2,\infty})$.

Im folgenden werden Beweistechniken aus der Konvergenztheorie der Finiten-Elemente (Ansatz- und Testraum identisch) auf die Konvergenztheorie des ASWR-Verfahrens (Ansatz- und Testraum verschieden) übertragen.

Dazu wird zunächst eine spezielle Norm eingeführt und deren Äquivalenz zur L_2 -Norm gezeigt.

3.3.2 Satz

Für $u, v \in \mathbf{D}_n$ gilt: $(u, v)_P := (Pu, Pv)_{L_2}$ ist ein Skalarprodukt.

und die Norm $\|u\|_P^2 := (u, u)_P$ läßt sich wie folgt abschätzen:

$$c_P \|u\|_{L_2} \leq \|u\|_P \leq \|u\|_{L_2}, \quad \forall u \in \mathbf{D}_n$$

Beweis:

P ist linear und damit auch $(\cdot, \cdot)_P$.

$$\begin{aligned} \text{Symmetrie:} \quad (u, v)_P &= (Pu, Pv)_{L_2} \\ &= (Pv, Pu)_{L_2} \\ &= (v, u)_P \end{aligned}$$

$$\begin{aligned} \text{Zudem gilt: } (Pu, Pv)_{L_2} &= (Pu, Pv)_{L_2} - (u - Pu, Pv)_{L_2} \quad \text{da } u - Pu \perp \mathbf{K}_n. \\ &= (u, Pv)_{L_2} \end{aligned}$$

$$(1) \quad = (Pu, v)_{L_2}$$

Positivität:

Sei $u \in \mathbf{D}_n$ mit $\|u\|_{L_2} = 1$
 $\|u\|_{L_2}^2 \leq c_{41} h^2 \|b\|_2^2$ mit $u = A_3 b$ siehe 2.1.19
 $\Rightarrow 1 \leq \sqrt{c_{41}} h \|b\|_2$

Sei weiter $v := b A_{\text{Ident}}$ mit $A_{\text{Ident}} = \text{Identitätsmatrix} \in \mathbb{R}^{n^2 \times \mathbb{R}^{n^2}}$
 $\Rightarrow v \in \mathbf{K}_n$.

$\Rightarrow \|v\|_{L_2} = h \|b\|_2$

$\Rightarrow (u, \frac{v}{\|v\|_{L_2}})_{L_2} = \frac{1}{\|v\|_{L_2}} b^T A_2 b$ mit A_2 aus 2.1.17

$\geq c_2 h \|b\|_2$ wegen 2.1.17

$\geq c_2 = c_p$ wegen $1 \leq \sqrt{c_{41}} h \|b\|_2$

$\Rightarrow \inf_{u \in \mathbf{D}_n} \sup_{v \in \mathbf{K}_n} (u, v)_{L_2} \geq c_p$
 (2) $\|u\|_{L_2} = 1 \quad \|v\|_{L_2} = 1$

Im weiteren seien $u, v \in \mathbf{D}_n$ so gewählt, daß $\|u\|_{L_2} = 1$ und $\|v\|_{L_2} = 1$ ist.

(3) $\|Pu\|_{L_2} \neq 0$, wegen $(u, v)_{L_2} \neq 0$ und (1)

Sei $e := \|Pu\|_{L_2} v - Pu$

$\Rightarrow \|Pu + e\|_{L_2} = \|Pu + \|Pu\|_{L_2} v - Pu\|_{L_2}$
 $= \|\|Pu\|_{L_2} v\|_{L_2}$

(4) $= \|Pu\|_{L_2} \|v\|_{L_2}$
 $= \|Pu\|_{L_2}$.

Weiter ist:

$(u, Pu + e)_{L_2} = \|Pu\|_{L_2} (u, v)_{L_2}$
 (5) $= c_p \|Pu\|_{L_2}$

Pu ist die beste L_2 -Approximation an u :

$\|Pu - u\|_{L_2}^2 \leq \|u - (Pu + e)\|_{L_2}^2$
 $= \|u\|_{L_2}^2 - 2(u, Pu + e)_{L_2} + \|Pu + e\|_{L_2}^2$
 (6) $\leq 1 - 2c_p \|Pu\|_{L_2} + \|Pu\|_{L_2}^2$ siehe (4) und (5).

Nun ist $(Pu, u - u)_{L_2} = 0 \Rightarrow$

$\|Pu\|_{L_2}^2 = \|u\|_{L_2}^2 - \|Pu - u\|_{L_2}^2$
 (7) $\leq 2c_p \|Pu\|_{L_2} - \|Pu\|_{L_2}^2$.

Da $\|Pu\|_{L_2}^2 \neq 0 \Rightarrow$

$$2 \|Pu\|_{L_2}^2 \leq 2 c_p \|Pu\|_{L_2} \Rightarrow$$

$$\|Pu\|_{L_2} \leq c_p \Rightarrow$$

$$\|Pu\|_{L_2} \leq c_p \|u\|_{L_2} \Rightarrow$$

$$\begin{aligned} (u, Pu)_{L_2} &= (Pu, Pu)_{L_2} \\ &\leq c_p^2 (u, u)_{L_2} . \end{aligned}$$

Mit (7) folgt:

$$\|Pu\|_{L_2}^2 \leq \|u\|_{L_2}^2$$

Damit ist Positivität gezeigt und aus (2) folgt die Behauptung



Der dämpfende Charakter des Zeitverlaufs der Gleichungen kommt in folgendem Lemma zum Vorschein:

3.3.3 Lemma

Seien $a \in \mathbf{C}[t_0, t_e]$ und $0 < m \in \mathbb{R}$.

Für $f \in \mathbf{C}^1[t_0, t_e]$ gelte:

$$(*) \quad \forall t \in [t_0, t_e] : \quad f(t)' \leq a(t) - m f(t).$$

Dann gilt:

$$\forall t \in [t_0, t_e] : \quad f(t) \leq e^{-m(t-t_0)} f(t_0) + \int_{t_0}^t e^{m(s-t)} a(s) \, ds.$$

Beweis:

$$\text{Ist} \quad f(t) \leq e^{-m(t-t_0)} f(t_0) + \int_{t_0}^t e^{m(s-t)} a(s) \, ds$$

$$\leq e^{-mt} (e^{mt_0} f(t_0) + \int_{t_0}^t e^{ms} a(s) \, ds).$$

Daraus folgt:

$$e^{mt} f(t) - \int_{t_0}^t e^{ms} a(s) \, ds \leq e^{mt_0} f(t_0).$$

Hat die linke Seite der Ungleichung fallenden Funktionsverlauf, so ist die Behauptung gezeigt.

Sei nun $z(t) := e^{mt} f(t) - \int_{t_0}^t e^{ms} a(s) ds$, $t \in [t_0, t_e]$.

$$z(t)' = e^{mt} (f(t)' + m f(t) - a(t)).$$

Mit (*) gilt $z(t)' \leq 0$ => Behauptung



Damit zum abschließenden Konvergenzsatz:

3.3.4 Satz

Sei $u_0 \in C^1([t_0, t_e], \mathbf{C}^*(\Omega))$ die Lösung des Anfangsrandwertproblems 3.1.2. Dann gilt für die ASWR-Lösung $u_A \in C^1([t_0, t_e], \mathbf{D}_n)$:

$$\begin{aligned} \|u_0(t) - u_A(t)\|_{L_2} &\leq \frac{1}{c_P} e^{-c_1(t-t_0)} \|g_A - Qg\|_P + \\ &\quad c_A h^2 \left(\|u_0(t)\|_* + \frac{1}{c_P} \int_{t_0}^t e^{c_1(s-t)} \left\| \frac{d}{dt} u_0(s) \right\|_* ds \right) \end{aligned}$$

Beweis:

Sei $u := Q u_0(t)$.

$u(t)$ ist also ASWR-Lösung von $\mathbf{L}u = \mathbf{L}u_0(t)$ mit entsprechenden Randbedingungen für $t \in [t_0, t_e]$. =>

$$(\mathbf{L}_\Sigma u, v)_{L_2} = (\mathbf{L}_\Sigma u_0, v)_{L_2}, \quad \forall v \in \mathbf{K}_n$$

$$B_\Sigma(u, v) = B_\Sigma(u_0, v), \quad \forall v \in \mathbf{K}_n.$$

Nach 3.2.1, der Definition der ASWR-Lösung, gilt:

$$\begin{aligned} \left(\frac{d}{dt} u_A, v \right)_{L_2} &= -B_\Sigma(u_A, v) + (r, v)_{L_2} \\ &= -B_\Sigma(u_A, v) + \left(\frac{d}{dt} u_0, v \right)_{L_2} + B_\Sigma(u_0, v) \\ &= B_\Sigma(u - u_A, v) + \left(\frac{d}{dt} u_0, v \right)_{L_2} \quad \Rightarrow \end{aligned}$$

$$\left(\frac{d}{dt} u_0 - \frac{d}{dt} u, v \right)_{L_2} = \left(\frac{d}{dt} u_A - \frac{d}{dt} u, v \right)_{L_2} + B_\Sigma(u_A - u, v), \quad \forall v \in \mathbf{K}_n$$

$$P(u_A - u) \in \mathbf{K}_n \quad \Rightarrow$$

(1)

$$\left(\frac{d}{dt} u_0 - \frac{d}{dt} u, P(u_A - u) \right)_{L_2} \geq \left(\frac{d}{dt} u_A - \frac{d}{dt} u, P(u_A - u) \right)_{L_2} + c_1 \|u_A - u\|_{L_2}^2,$$

wegen 3.1.6.

$$\begin{aligned}
\|u_A - u\|_P \cdot \frac{d}{dt} \|u_A - u\|_P &= \frac{1}{2} \frac{d}{dt} (\|u_A - u\|_P^2) \\
&= \frac{1}{2} \frac{d}{dt} (u_A - u, P(u_A - u))_{L_2} \\
&= \frac{1}{2} \left(\left(\frac{d}{dt} (u_A - u), P(u_A - u) \right)_{L_2} + (u_A - u, \frac{d}{dt} P(u_A - u))_{L_2} \right) \\
&= \left(\frac{d}{dt} (u_A - u), P(u_A - u) \right)_{L_2}.
\end{aligned}$$

$$\begin{aligned}
\text{Da } \|P(u_A - u)\|_{L_2}^2 &= (P(u_A - u), P(u_A - u))_{L_2} \\
&= (u_A - u, P(u_A - u))_{L_2} \\
&= \|u_A - u\|_P^2.
\end{aligned}$$

Mit (1) und der Schwarz'schen Ungleichung $\Rightarrow$

$$\left\| \frac{d}{dt} u_0 - \frac{d}{dt} u \right\|_{L_2} \|u_A - u\|_P \geq \|u_A - u\|_P \frac{d}{dt} \|u_A - u\|_P + c_1 \|u_A - u\|_P^2$$

$$\text{Aus } \|u_A - u\|_P \neq 0 \quad \Rightarrow$$

$$\left\| \frac{d}{dt} u_0 - \frac{d}{dt} u \right\|_{L_2} \geq \frac{d}{dt} \|u_A - u\|_P + c_1 \|u_A - u\|_P.$$

Aus 3.3.3 folgt:

$$\begin{aligned}
\|u_A(t) - u(t)\|_P &\leq e^{-c_1(t-t_0)} \|u_A(0) - u(0)\|_P + \\
&\quad \int_{t_0}^t e^{c_1(s-t)} \left\| \frac{d}{dt} u_0(s) - \frac{d}{dt} u(s) \right\|_{L_2} ds.
\end{aligned}$$

Schließlich gilt mit der Dreiecksungleichung und obiger Aussage:

$$\begin{aligned}
\|u_A(t) - u_0(t)\|_{L_2} &\leq \|u_0(t) - u(t)\|_{L_2} + \|u_A(t) - u(t)\|_{L_2} \\
&\leq \|u_0(t) - u(t)\|_{L_2} + \frac{1}{c_P} \|u_A(t) - u(t)\|_P \quad \text{siehe 3.3.2} \\
&\leq \|u_0(t) - Q u_0(t)\|_{L_2} + \frac{1}{c_P} (e^{-c_1(t-t_0)} \|g_A - Q g\|_P + \\
&\quad \int_{t_0}^t e^{c_1(s-t)} \left\| \frac{d}{dt} u_0(s) - \frac{d}{dt} Q u_0(s) \right\|_{L_2} ds)
\end{aligned}$$

mit Eingang von 2.3.6 gilt weiter

$$\begin{aligned}
\|u_A(t) - u_0(t)\|_{L_2} &\leq c_{A1} h^2 \|u_0(t)\|_* + \frac{1}{c_P} (e^{-c_1(t-t_0)} \|g_A - Q g\|_P + \\
&\quad c_{A1} h^2 \int_{t_0}^t e^{c_1(s-t)} \left\| \frac{d}{dt} u_0(s) \right\|_* ds)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{c_p} e^{-c_1(t-t_0)} \|g_A - Qg\|_p + \\
&\quad c_A h^2 \left(\|u_0(t)\|_* + \frac{1}{c_p} \int_{t_0}^t e^{c_1(s-t)} \left\| \frac{d}{dt} u_0(s) \right\|_* ds \right)
\end{aligned}$$

□

$\frac{1}{c_p} e^{-c_1(t-t_0)} \|g_A - Qg\|_p$ beschreibt den Einfluß der Approximation der Anfangsbedingung. Wählt man als Anfangsapproximation $g_A = Qg$, so fällt dieser Term weg. g_A ist somit ein günstigerer Startwert.

4 Entwicklung eines Verfahrens vom ASWR-Typ für die partiellen Differentialgleichungen der Festkörperdiffusion

Nachdem nun die mathematische Analyse die Konvergenz des ASWR-Verfahrens in der L_2 -Norm gezeigt hat, werden nun algorithmische Gesichtspunkte der Methode in ihrer praktischen Anwendung auf die partiellen Differentialgleichungen der Festkörperdiffusion betrachtet. Hierzu ist unter anderem ein tieferes Verständnis des physikalischen Hintergrunds notwendig.

Seitdem es möglich ist, auf einem Chip Einzelbauteile zu integrieren, steigt auch die Anzahl der auf einem Chip integrierten Einzeltransistoren. Mit der fortschreitenden Miniaturisierung der Einzeltransistoren mit Kanallängen von einem μm ist die Prozeßtechnik derart kompliziert und aufwendig geworden, daß eine Simulation der Prozeßschritte viel Zeit und Geld sparen kann. Zudem werden mit abnehmenden Strukturgrößen die Dotierungskonzentrationen höher und die Prozeßtemperatur niedriger, was die Verwendbarkeit von analytischen, linearen Modellen ausschließt.

Diese Arbeit beschäftigt sich nun allein mit der Simulation der Diffusion von dotierten Fremdatomen im Silicium. Die ins Silicium hineingeschossenen oder hineindiffundierten Fremdatome werden aufgrund hoher Temperaturen (700 - 1200 Celsius) umverteilt. Diese Umverteilung im Silicium wird durch ein System von zeitabhängigen partiellen Differentialgleichungen beschrieben. Die Gleichungen lassen sich nur in den einfachsten Fällen analytisch lösen, sodaß für die Lösung numerische Verfahren benutzt werden.

4.1 Das System partieller Differentialgleichungen zur Beschreibung der Festkörperdiffusion

Als erstes möchte ich vorausschicken, daß speziell bei der Miniaturisierung der Transistoren durch hohe Konzentrationen der Fremdatome Effekte auftreten, die meist nur phänomenologisch erfaßt sind. So stützt sich die eigentliche Modellbildung auf Voraussetzungen, die besonders im hochkonzentrierten Fall nur sehr beschränkt erfüllt sind. Für diesen Fall versucht man das Modell so zu erweitern, die Parameter so einzustellen, daß die Simulation mit gemessenen Ergebnissen übereinstimmt.

Das Diffusionsmodell begründet sich in dem modifizierten Fick'schen Gesetz, das besagt, daß der Diffusionsstrom j proportional zum Gradient des chemischen Potentials μ ist. C beschreibt die Fremdatomkonzentration und u beschreibt die Beweglichkeit der Teilchen:

$$(1) \quad j = -C u \text{ grad } \mu.$$

Man setzt für das chemische Potential μ folgende Gleichung an

$$(2) \quad \mu = \mu_0 + kT \ln(C) + zq\Phi,$$

wobei zq die elektrische Ladung des Dotierstoffes,
 Φ das elektrostatische Potential,
 k Boltzmann Konstante,
 μ_0 Konstante und
 T Temperatur ist.

Berücksichtigt man bei der Darstellung des chemischen Potentials μ die Ladung des elektrostatischen Potentials nicht, so erhält man aus der Darstellung (1) das klassische Fick'sche Gesetz

$$j = -D \text{ grad } C$$

mit D als Proportionalitätskonstante zum Diffusionsstrom j . Damit erhält man für die zeitliche Änderung:

$$(3) \quad \frac{\partial C}{\partial t} = -\text{div } j = \text{div } D \text{ grad } C.$$

Diese Gleichung gilt jedoch nur für niedrige Konzentrationen. Eine Diffusionsverstärkung für hohe Konzentrationen ist mit (3) nicht möglich. Der Verstärkungseffekt wird jedoch durch die Einbeziehung des elektrostatischen Potentials Φ und durch das nichtlineare Verhalten des Diffusionskoeffizienten D erreicht. Die physikalische Erklärung des nichtlinearen Verhaltens von D beruht auf der Annahme, daß aufgrund der etwa gleichen Größe der Fremdatome zu den Siliciumatomen die Fremdatome nur über geladene Leerstellen im Silicium diffundieren können. Als Maß für die Anzahl der geladenen Leerstellen wird die Anzahl der beweglichen Elektronen genommen, die wiederum abhängig von den Konzentrationen der Fremdatome sind. Die Konzentration nicht festgebundener Elektronen bzw. Löcher beeinflußt die Ladung und die Konzentration der verschiedenen geladenen Leerstellen. Vereinfachend nimmt man an, daß Gruppe V-Elemente wie Phosphor und Arsen teilweise über negativ geladene Leerstellen, Gruppe III-Elemente wie Bor teilweise über positive Leerstellen diffundieren. Man berücksichtigt diesen Einfluß durch Einführung additiver Terme zum idealen, allein temperaturabhängigen Diffusionskoeffizienten. Dabei wird die Elektronendichte bzw. Löcherdichte durch die im dotierten Fall aufgrund von thermischer Anregung im Silicium vorhandenen Elektronen bzw. Löcherkonzentration n_i bzw. p_i dividiert. So erhält man einen Diffusionskoeffizienten, der nichtlinear abhängig ist von den Fremdatomkonzentrationen. Damit kommt man zur folgenden Darstellung des Diffusionskoeffizienten:

$$(4a) \quad D_D = D_i + D^- \frac{n}{n_i},$$

$$(4b) \quad D_A = D_i + D^+ \frac{p}{p_i}.$$

(4a) gilt für negative Ladungsträger und (4b) für positive. n bzw. p bezeichnen die Elektronen- bzw. Löcherkonzentrationen (ein Loch entspricht einem

fehlenden Elektronen). n_i ist die Eigenleitdichte und beschreibt die im reinem Silicium freien Elektronen, die nur abhängig von der Temperatur T ist. Daher gilt nach dem Massenwirkungsgesetz:

$$(5) \quad n p = n_i^2 \approx \text{const} \exp(-E_g/2kT).$$

Die Elektronenkonzentration berechnet sich aus der Boltzmann-Statistik und ist abhängig von der Eigenleitdichte n_i und den Konzentrationen der Fremdatome. Je nachdem, ob die Fremdatome Elektronen oder Löcher in den Kristall bringen, gehen diese mit positivem oder negativem Vorzeichen in die Statistik ein. C_D (D ist Abkürzung für Donatoren und A für Akzeptoren) bezeichnet hierbei die Fremdatomkonzentrationen, die negative Ladungen transportieren; C_A ist positiver Ladungsträger. n läßt sich folgendermaßen darstellen:

$$(6) \quad n = \frac{\sum C_D - \sum C_A + \sqrt{(\sum C_D - \sum C_A)^2 + 4n_i^2}}{2}.$$

Sind keine Fremdatome vorhanden, so ist logischerweise $n = n_i$ was auch in obiger Formel zur Geltung kommt. Damit ist auch die Abhängigkeit des Diffusionskoeffizienten von den Fremdatomkonzentrationen klar. So hängt der Diffusionskoeffizient eines Dopanden nicht nur von der eigenen Konzentration ab, sondern auch von den Konzentrationen der anderen Dopanden. Damit ist eine Kopplung der Dopandengleichungen schon ersichtlich. Eine weitere ist durch den Feldterm gegeben, also die Berücksichtigung des elektrostatischen Potentials Φ in den Gleichungen. Zunächst läßt sich der Zusammenhang zwischen den Dopandenkonzentrationen und dem elektrostatischen Potential durch die Poisson-Gleichung beschreiben:

$$(7) \quad 2n_i \sinh\left(\frac{z q \Phi}{kT}\right) - k \operatorname{div} \operatorname{grad} \Phi = C \quad \text{mit } k = \text{Konstante.}$$

Diese Differentialgleichung läßt sich durch Weglassen des Differentialterms auf eine algebraische Gleichung reduzieren (quasineutraler Ansatz). Baut man dies wiederum in (2) und (1) ein so erhält man:

$$(8a) \quad \frac{\partial C_D}{\partial t} = \operatorname{div} \left(D_D (\operatorname{grad} C_D + C_D \operatorname{grad} \Phi) \right),$$

$$(8b) \quad \frac{\partial C_A}{\partial t} = \operatorname{div} \left(D_A (\operatorname{grad} C_A - C_A \operatorname{grad} \Phi) \right) \quad \text{mit}$$

$$\Phi = \ln\left(\frac{n}{n_i}\right) \quad \text{aus dem quasineutralen Ansatz von (7).}$$

Dabei wurde der Nullpunkt des Potentials in das Fermi-Niveau gelegt, also Φ ist für $n = n_i$ Null. Zusammen mit (6) und (5) sind nun die Gleichungen mit C_D und C_A als unabhängige Variablen gegeben. Somit hängt die Anzahl der Gleichungen von der Anzahl der Dopanden ab, wobei für Akzeptoren die

Gleichung (8b) und für Donatoren (8a) gilt. Unberücksichtigt bleibt in der Darstellung die Bildung von Komplexen (Cluster), also das Zusammenschließen von zwei Fremdatomen oder mehreren zu einem Molekül: Bei sehr hohen Konzentrationen (z.B. Arsen 5.0×10^{20} Atome pro cm^3 bei 1000°) schwächt sich die Diffusion ab. Das führt man auf den Effekt zurück, daß diese Fremdatommoleküle aufgrund ihrer Größe und festen Bindung nicht diffundieren können. Dieser Effekt korrespondiert mit einer Art Sättigungsgrenze der elektrisch aktiven Atome, die durch die Molekülbildung erklärt wird. Das heißt, werden mehr Fremdatome in das Silicium hineingeschossen, so erhöht sich logischerweise auch die Wahrscheinlichkeit, daß zwei benachbarte Atome Fremdatome sind, somit ein Molekül bilden und also auch kein Ladungsträger für den Stromtransport sind.

4.2 Probleme bei der Anwendung numerischer Methoden

Die Abhängigkeit des elektrostatischen Potentials und der totalen Elektronenkonzentration von den unabhängigen Variablen C ist, wie oben ersichtlich, nichtlinear. Damit ist auch die Kopplung des Gleichungssystems nichtlinear, so daß bei impliziter Zeitdiskretisierung auch das Gleichungssystem nichtlinear ist. Implizite Diskretisierung der in der Zeit "steifen" Gleichungen ist notwendig, da ansonsten mit einem instabilen numerischen Verhalten gerechnet werden muß. Um den Diskretisierungsfehler möglichst klein zu halten, wird das Gitter den physikalischen Gegebenheiten angepaßt, wie z.B. durch die Berücksichtigung von hohen zweiten und dritten Ableitungen im elektrostatischen Potential.

Unangenehm ist jedoch dabei, daß diese Stellen nicht über den Zeitverlauf an einem Ort festliegen, sondern sich mit der Zeit verlagern. Um diesem Rechnung zu tragen, ist ein zeitabhängiges Gitter notwendig, das sich selbst, den Gegebenheiten folgend, steuert.

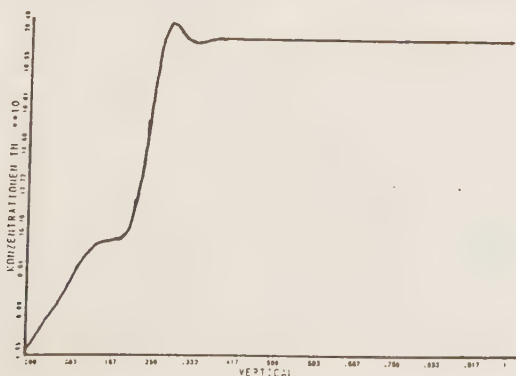
4.3 Motivation für die Untersuchung von Verfahren des ASWR-Typs für die Festkörperdiffusion

Die ersten Simulationsprogramme der Festkörperdiffusion wurden mit Finiten-Differenzen geschrieben, die man wegen ihrer einfachen und relativ schnellen Programmierung bevorzugte. Um jedoch eine entsprechende Rechengenauigkeit in zwei Raumdimensionen zu erhalten, muß mit einem entsprechend feinen Gitter gearbeitet werden, so daß die Rechenzeiten für hochkonzentrierte Dopanden selbst auf schnellen Rechnern (wie CDC 175) in den Stundenbereich stoßen. Der Kosten- und Zeitvorteil, der gerade durch die Simulation eintreten soll, wäre somit nicht mehr signifikant gegeben. Dieser Punkt soll mit dem ASWR-Verfahren verbessert werden, um womöglich auf lange Sicht auch dreidimensionale Geometrien behandeln zu können. Tatsächlich zeigen auch die bisherigen Erfahrungen eine deutliche Verminderung der Rechenzeit bis zum Faktor 10 und mehr. Somit ist eine Anwendung für den dreidimensionalen Fall durchaus denkbar. Der direkte Vergleich der Methoden ist jedoch schwierig, da für die ASWR-Methode eine der Problematik entsprechende Transformation der Gleichungen vorgenommen wird, die durchaus auch bei der Anwendung der Finiten-Differenzen zu einer Rechenbeschleunigung hätte beitragen können.

4.4 Transformation der Differentialgleichungen

Bei den ersten Versuchen, das ASWR-Verfahren für die Simulation der Festkörperdiffusion zu verwenden, traten Probleme auf, wie man sie auch bei den Finiten-Elementen mit höhergradigen Polynomfunktionen her kennt. Wegen der großen Gradienten (10^6) und der großen zweiten Ableitungen beginnen die Ansatzfunktionen zu "oszillieren", das heißt, bei starker Variation des Gradienten approximieren die Ansatzfunktionen den tatsächlichen Verlauf qualitativ falsch.

Beispiel:



Um dieses Problem zu entschärfen, wird eine Transformation der Differentialgleichungen durchgeführt. Dabei geht man von folgender Idee aus: Die Anfangsverteilungen bzw. die Anfangsbedingungen der Gleichungen werden durch Gaussverteilungen bestimmt, wie als Beispiel folgende eindimensionale Verteilung:

$$C(x) = \frac{10^{20}}{s} e^{-\left(\frac{x - x_m}{s}\right)^2} \quad \text{mit}$$

x_m Mittelwert und
 s Streuung der Verteilung.

Um nun in der numerischen Darstellung große Werte und Gradienten zu vermeiden, soll nicht die ganze Verteilungsfunktion dargestellt werden, sondern nur der Exponent der Funktion. Da der Exponent in diesem Fall nur eine quadratische Funktion ist, können also hier die Ansatzfunktionen des ASWR-Verfahrens diesen exakt darstellen und somit auch die ganze Verteilung. Was also im folgenden gezeigt wird, ist die Darstellung des Exponenten in den Differentialgleichungen:

Der Ansatz $C(x,t) = e^{f(x,t)}$

liefert

$$\begin{aligned} \frac{\partial}{\partial x} \left(D \frac{\partial}{\partial x} C \right) &= \frac{\partial}{\partial x} \left(D \frac{\partial}{\partial x} e^f \right) \\ &= \frac{\partial}{\partial x} \left(D e^f \frac{\partial f}{\partial x} \right) \\ &= e^f \left(\frac{\partial}{\partial x} \left(D \frac{\partial f}{\partial x} \right) + D \left(\frac{\partial f}{\partial x} \right)^2 \right). \end{aligned}$$

Damit erscheint im Differentialterm nur noch f . Analog gilt für den Feldterm:

$$\frac{\partial}{\partial x} \left(D e^f \frac{\partial \Phi}{\partial x} \right) = e^f \left(\frac{\partial}{\partial x} \left(D \frac{\partial \Phi}{\partial x} \right) + D \left(\frac{\partial f}{\partial x} \right) \left(\frac{\partial \Phi}{\partial x} \right) \right).$$

Ohne das Vorkommen von hohen Konzentrationen läßt sich Gleichung 8a in 4.1 mit Hilfe von f darstellen. Für eine Raumdimension gilt somit:

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial x} D \left(\frac{\partial f}{\partial x} + \frac{\partial \Phi}{\partial x} \right) + D \frac{\partial f}{\partial x} \left(\frac{\partial f}{\partial x} + \frac{\partial \Phi}{\partial x} \right).$$

Analog behandelt man den zweidimensionalen Fall. Ein weiterer Vorteil dieser Darstellung ist der in etwa gleiche Wertebereich von Φ und f . So liegen im Vergleich dazu die Konzentrationen zwischen 0 und 10^{21} cm^{-3} (wobei aber schon 10^{10} cm^{-3} als signifikant 0 angesehen werden kann), während Φ zwischen -5 und 15 und f zwischen 0 und 50 liegen. Nachteil dieser

Darstellung ist eine weitere Nichtlinearität in Form eines Produkts zweier Ableitungen im hinzugekommenen Differentialterm, die sich jedoch in der Anwendung als nicht relevant erweist. So ist es möglich, die obige Gleichung ohne Feldterme und mit konstantem Diffusionskoeffizienten D , mit einer Gaussverteilung als Anfangsverteilung und unter Weglassung des Zeitdiskretisierungsfehlers (Berechnung mit sehr kleinen Zeitschritten) mit wenigen Gitterlinien exakt zu berechnen.

4.5 Semidiskrete ASWR-Lösung für die Halbleitergleichungen

In Kapitel 2 ist die ASWR-Lösung für ein Randwertproblem definiert worden. Die Halbleitergleichungen sind, wie oben gezeigt, transient und müssen somit zur numerischen Handhabung auch in der Zeit diskretisiert werden. Zuvor soll jedoch für das Problem die semi-diskrete ASWR-Lösung eingeführt werden, sodaß ein System von gewöhnlichen Differentialgleichungen in der Zeit zu lösen bleibt. Die Funktionsauswertungen der gewöhnlichen Differentialgleichungen sind die mit der ASWR-Methode diskretisierten partiellen räumlichen Ableitungen. Weiter ist zu bemerken, daß die nun folgende ASWR-Definition nicht wie in Kapitel 2 als Variationsgleichung formuliert ist, sondern als Integralgleichung. Tatsächlich ist unsere Formulierung mit der in Kapitel 2 äquivalent; in der folgenden Form wird die spezielle Art der Testfunktionen (stückweise Konstante) in die Bilinearform eingesetzt:

4.5.1 Definition

$$\begin{aligned} \frac{\partial}{\partial t} \int_{\Omega_{ik}} f_A d(x,y) &= \int_{\Omega_{ik}} \frac{\partial}{\partial x} D_A \left(\frac{\partial f_A}{\partial x} + \frac{\partial \Phi_A}{\partial x} \right) + D_A \frac{\partial f_A}{\partial x} \left(\frac{\partial f_A}{\partial x} + \frac{\partial \Phi_A}{\partial x} \right) d(x,y) + \\ &\quad \int_{\Omega_{ik}} \frac{\partial}{\partial y} D_A \left(\frac{\partial f_A}{\partial y} + \frac{\partial \Phi_A}{\partial y} \right) + D_A \frac{\partial f_A}{\partial y} \left(\frac{\partial f_A}{\partial y} + \frac{\partial \Phi_A}{\partial y} \right) d(x,y) \end{aligned}$$

für alle Ω_{ik} (definiert in 2.2.2) aus Ω

mit $f_A \in \mathbf{D}_n$, $\Phi_A := \Phi(f_A)$ und $D_A := D(f_A)$.

4.6 Zeitdiskretisierung und Nichtlinearitäten

Wie oben beschrieben, werden mit der ASWR-Methode die räumlichen Ableitungen diskretisiert. Übrig bleiben die zeitlichen, die ein System von gewöhnlichen Differentialgleichungen bilden und deren Dimension von der Anzahl der Ω_{ik} und der Anzahl der Dopanden abhängt.

Prinzipiell sind Verfahren zur Lösung von gewöhnlichen Differentialgleichungen in explizite und implizite Methoden unterscheidbar. Explizite Verfahren sind programmtechnisch leicht zu verifizieren, führen jedoch zu Stabilitätsproblemen [GRI-72] bei sogenannten "steifen Problemen". Implizite Verfahren sind meist absolut stabil, d.h. obige Probleme treten nicht auf, sind aber in der Programmrealisierung schwieriger und aufwendiger zu handhaben, da bei mehreren gekoppelten Gleichungen Gleichungssysteme zu lösen sind. Für die klassische Anwendung der Trapezregel [HÄM-72] müßte so für obige Gleichungen ein nichtlineares Gleichungssystem gelöst werden.

$$\begin{aligned}
 4.6.1 \quad \int_{\Omega_{ik}} f_A^{j+1} d(x,y) - \int_{\Omega_{ik}} f_A^j d(x,y) = \\
 \frac{1}{2} \Delta t^j \left(\int_{\Omega_{ik}} \frac{\partial}{\partial x} D_A^{j+1} \left(\frac{\partial f_A^{j+1}}{\partial x} + \frac{\partial \Phi_A^{j+1}}{\partial x} \right) + D_A^{j+1} \left(\frac{\partial f_A^{j+1}}{\partial x} \left(\frac{\partial f_A^{j+1}}{\partial x} + \frac{\partial \Phi_A^{j+1}}{\partial x} \right) \right) d(x,y) + \right. \\
 \left. \int_{\Omega_{ik}} \frac{\partial}{\partial y} D_A^{j+1} \left(\frac{\partial f_A^{j+1}}{\partial y} + \frac{\partial \Phi_A^{j+1}}{\partial y} \right) + D_A^{j+1} \left(\frac{\partial f_A^{j+1}}{\partial y} \left(\frac{\partial f_A^{j+1}}{\partial y} + \frac{\partial \Phi_A^{j+1}}{\partial y} \right) \right) d(x,y) + \right. \\
 \left. \int_{\Omega_{ik}} \frac{\partial}{\partial x} D_A^j \left(\frac{\partial f_A^j}{\partial x} + \frac{\partial \Phi_A^j}{\partial x} \right) + D_A^j \frac{\partial f_A^j}{\partial x} \left(\frac{\partial f_A^j}{\partial x} + \frac{\partial \Phi_A^j}{\partial x} \right) d(x,y) + \right. \\
 \left. \int_{\Omega_{ik}} \frac{\partial}{\partial y} D_A^j \left(\frac{\partial f_A^j}{\partial y} + \frac{\partial \Phi_A^j}{\partial y} \right) + D_A^j \frac{\partial f_A^j}{\partial y} \left(\frac{\partial f_A^j}{\partial y} + \frac{\partial \Phi_A^j}{\partial y} \right) d(x,y) \right)
 \end{aligned}$$

wobei f_A^j Werte nach j , f_A^{j+1} Werte nach $j+1$ Zeitschritten sind. Δt^j bezeichnet die Zeitdifferenz von $j+1$ und j Zeitschritten.

Das Gleichungssystem bestimmt aus der bekannten Schicht j die neue Schicht $j+1$. Da in der Schicht $j+1$ auch D und Φ , zudem der Differentialoperator nichtlinear sind, ist auch das aus obiger Diskretisierung resultierende Gleichungssystem nichtlinear. Mit (4.6.1) ist also ein nichtlineares Gleichungssystem zu lösen, was nur mit rechenintensiven Iterationen zu bewerkstelligen ist.

Um Letzteres zu vermeiden, werden in der nächsten Darstellung die "Nichtlinearitäten explizit" dargestellt, sodaß nur ein lineares Gleichungssystem zur Bestimmung der neuen Zeitschicht gelöst werden muß. Für die implizite Darstellung des elektrostatischen Potentials Φ wird eine Linearisierung verwendet, die eine "relativ implizite" Darstellung ermöglicht. Die Linearisierung erfolgt durch die Kettenregel:

$$4.6.2 \quad \frac{\partial \Phi_A^{j+1}}{\partial x} := \frac{\partial \Phi_A^j}{\partial n^j} \frac{\partial n^j}{\partial C_A} \frac{\partial C_A^{j+1}}{\partial x}.$$

$\frac{\partial \Phi_A^j}{\partial n^j}$ und $\frac{\partial n^j}{\partial C_A}$ werden als bekannt vorausgesetzt, da sie aus der Schicht j stammen. Damit werden die unbekannten partiellen Ableitungen in Schicht $j+1$ durch die partiellen Ableitungen von C ersetzt, die ja linear in das Gleichungssystem eingehen. Bei Testrechnungen erwies sich die Linearisierung von (4.6.2) als sehr brauchbar (bei adequaten Zeitschrittweiten). Für die Diffusionskoeffizienten und dem quadratischen Differentialterm kann aufgrund praktischer Ergebnisse auf eine Linearisierung verzichtet werden. Dort reicht eine explizite Darstellung. Konkret erfolgt die Diskretisierung für die Zeitintegration in drei Zeitschichten, um aus der mittleren Schicht Diffusionskoeffizienten und quadratischen Differentialterm zu errechnen. Somit werden die nicht "implizit" genommenen Terme in einem Zeitschritt gemittelt:

$$4.6.3 \quad \int_{\Omega_{ik}} f_A^{j+2} d(x,y) - \int_{\Omega_{ik}} f_A^j d(x,y) =$$

$$(\Delta t^j + \Delta t^{j+1}) \left(\frac{1}{2} \left(\int_{\Omega_{ik}} \left(\frac{\partial}{\partial x} D_A^{j+1} \left(\frac{\partial f_A^{j+2}}{\partial x} + \frac{\partial \Phi_A^{j+2}}{\partial x} \right) + \frac{\partial}{\partial y} D_A^{j+1} \left(\frac{\partial f_A^{j+2}}{\partial y} + \frac{\partial \Phi_A^{j+2}}{\partial y} \right) \right) d(x,y) \right. \right.$$

$$+ \left. \int_{\Omega_{ik}} D_A^{j+1} \left(\frac{\partial f_A^j}{\partial x} + \frac{\partial \Phi_A^j}{\partial x} \right) + \frac{\partial}{\partial y} D_A^{j+1} \left(\frac{\partial f_A^j}{\partial y} + \frac{\partial \Phi_A^j}{\partial y} \right) d(x,y) \right) +$$

$$\int_{\Omega_{ik}} D_A^{j+1} \left(\left(\frac{\partial f_A^{j+1}}{\partial x} \left(\frac{\partial f_A^{j+1}}{\partial x} + \frac{\partial \Phi_A^{j+1}}{\partial x} \right) + \left(\frac{\partial f_A^{j+1}}{\partial y} \left(\frac{\partial f_A^{j+1}}{\partial y} + \frac{\partial \Phi_A^{j+1}}{\partial y} \right) \right) d(x,y) \right).$$

Aus Schicht j und $j+1$ werden die Werte aus Schicht $j+2$ berechnet. In dieser Form ist das Gleichungssystem linear. Hat man nur eine bekannte Zeitschicht zur Verfügung, wie z.B. zu Beginn oder bei Einführung eines neuen Gitters, so setzt man $j := j+1$ und $\Delta t^j := 0$.

4.6.4 Bemerkung

Durch die Transformation der Differentialgleichung in (4.2) hat sich auch das Verhalten der Zeitapproximation verändert. Die Zeitgradienten sind für hohe Konzentrationen in der transformierten Gleichung nicht so groß, was zur Folge hat, daß die Approximationsgüte bei gleichen Verfahren zumindest bei hohen Konzentrationen verbessert ist. Auch wenn das für niedrige Konzentrationen nicht gilt, kann man sagen, daß die Zeitapproximation aufgrund der Transformation gutmütiger geworden ist.

4.7 Orts - und zeitabhängiger Gittergenerator

Ein nichtäquidistantes Gitter wird, wenn es sich der Lösung anpaßt, Rechenzeit und Speicherplatz sparen helfen. Der Fehler der Ortsdiskretisierung wird durch die Approximationsgüte der quadratischen Splines an die Lösung bestimmt, was auch aus den Abschätzungen im Kapitel 1 und 2 zu ersehen ist. Also wird das Gitter so zu legen sein, daß der Approximationsfehler der Splines möglichst klein wird.

Die Lösungen ändern sich während des zeitlichen Ablaufs signifikant. Es ist deswegen ratsam, das Gitter den jeweiligen neuen Umständen anzupassen. Das führt zur Problematik des zeitabhängigen Gitters. Der Benutzer hat während der Simulation keine Eingriffsmöglichkeit in die Numerik, sodaß ein Gittergenerator selbständig entscheiden muß, wann und wie ein neues Gitter zu bilden ist. Dabei soll ein allzu häufiges Umgenerieren des Gitters vermieden werden, weil sonst die Kosten für das Neuanlegen des Gitters im Vergleich zu den eigentlichen Berechnungen zu groß und somit unrentabel werden.

Das folgende Resultat von Pereyra und Sewell [PER-75] gibt einen Hinweis, wie der Interpolationsfehler von stückweisen Polynomen zu minimieren ist:

4.7.1 Lemma

Sei $\mathbf{x}$ (definiert in 1.2) eine Unterteilung auf dem Intervall $[a,b]$ in n Intervalle und sei u die zu interpolierende Funktion aus $\mathbf{C}^1$. Das stückweise Polynom aus $\mathbf{P}_{1,\mathbf{x}}$ vom Grad 1 auf den Intervallen $(x_{i-1/2}, x_{i+1/2})$, $i=1, \dots, n$, das u auf $\mathbf{x}$ interpoliert, hat einen minimalen L_2 -Fehler, falls $\mathbf{x}$ so gewählt ist, daß

$$h_i^{1+1} |u^{(1+1)}(x_i)| = \text{const} \quad , \quad \forall i = 1, \dots, n.$$

Das Lemma zeigt, daß der Interpolationsfehler minimal wird, falls $h_i^{l+1} |u^{(l+1)}(x_i)|$ gleichverteilt ist. Da man die exakte Lösung u nicht kennt, wird diese durch die auf dem alten Gitter approximierte Lösung $\bar{u}$ ersetzt. $\bar{u}$ ist aus $\mathbf{P}_2, \mathbf{x}$, also quadratisches Polynom; somit ist $\bar{u}^{(3)} = 0$. Differenzenbildung ist ein Ausweg; man approximiert $u^{(3)}$ durch die Differenz der zweiten Ableitungen in benachbarten Intervallen. Die Näherung von $u^{(3)}$ ist offensichtlich nicht sehr genau, sodaß der Algorithmus, der die Gleichverteilung mit Hilfe von $\bar{u}^{(3)}$ berechnet, nicht sehr sensitiv von diesem Wert abhängen darf, da sonst der Diskretisierungsfehler von $\bar{u}^{(3)}$ das neue Gitter zu stark beeinflusst. Sinnvoll ist es, auf ein vollkommen neues Gitter, das aufgrund von $u^{(3)}$ berechnet würde, zu verzichten, und nur dort das alte Gitter zu ändern, wo signifikant Änderungen notwendig sind. So soll z.B. nur dort der Gitterabstand halbiert werden, in dessen Intervall der Wert $h_i^3 \bar{u}^{(3)}$ um ein Vielfaches über dem Durchschnittswert liegt. Der Faktor, ab dem ein Gitterwechsel erfolgt, steuert auch die Häufigkeit des Gitterwechsels, der ja, wie oben schon erwähnt, nicht zu oft stattfinden soll. Akzeptable Werte für diesen Faktor liegen zwischen 2.5 und 6.

4.8 Anpassung des Gitters für die Zeitdiskretisierung

Obige Betrachtungen zielen auf die Minimierung des Ortsdiskretisierungsfehler. Wichtiger, zumindest bei der hier behandelten Thematik, ist es ein Gitter zu erzeugen, das eine günstige Zeitdiskretisierung ermöglicht. Der Einfluß des Gitters auf die Lösung der zeitabhängigen Gleichungen läßt sich am Diskretisierungsfehler der Trapezregel und an der Steifigkeit des Systems beobachten. Im folgenden der Diskretisierungsfehler $\tau_{\Delta t}$ [GRI-72] für die Trapezregel:

$$\begin{aligned}
 4.8.1 \quad \tau_{\Delta t}(t) &= \frac{\Delta t^2}{6} \left(-\frac{1}{2} \left(\frac{\partial^2}{\partial t^2} \bar{F} + 2\bar{F} \frac{\partial}{\partial t} \frac{\partial}{\partial \bar{u}} \bar{F} + \bar{F}^2 \frac{\partial^2}{\partial \bar{u}^2} \bar{F} \right) + \right. \\
 &\quad \left. \frac{\partial}{\partial t} \bar{F} \frac{\partial}{\partial \bar{u}} \bar{F} + \bar{F} \frac{\partial^2}{\partial \bar{u}^2} \bar{F} \right) \quad \text{mit} \\
 \bar{u} &= \int_{\Omega_{ik}} f d(x,y) \quad \text{und} \\
 \bar{F} &= \int_{\Omega_{ik}} \left(\frac{\partial}{\partial x} D \left(-\frac{\partial f}{\partial x} + \frac{\partial \Phi}{\partial x} \right) + \frac{\partial}{\partial y} D \left(-\frac{\partial f}{\partial y} + \frac{\partial \Phi}{\partial y} \right) \right) d(x,y) + \\
 &\quad \int_{\Omega_{ik}} D \left(\left(\frac{\partial f}{\partial x} \left(-\frac{\partial f}{\partial x} + \frac{\partial \Phi}{\partial x} \right) + \left(\frac{\partial f}{\partial y} \left(-\frac{\partial f}{\partial y} + \frac{\partial \Phi}{\partial y} \right) \right) \right) d(x,y).
 \end{aligned}$$

Approximiert man $\bar{F}$ durch $h_i h_k F(x_i, y_k)$ und $\bar{u}$ durch $h_i h_k u(x_i, y_k)$, so verhält sich

$$\frac{\partial^2}{\partial t^2} \bar{F} \quad \text{wie} \quad h_i h_k \frac{\partial^2}{\partial t^2} F(x_i, y_k),$$

$$\frac{\partial}{\partial t} \frac{\partial}{\partial \bar{u}} \bar{F} \quad \text{wie} \quad \frac{\partial}{\partial t} \frac{\partial}{\partial u} F(x_i, y_k),$$

$$\frac{\partial^2}{\partial \bar{u}^2} \bar{F} \quad \text{wie} \quad \frac{1}{h_i h_k} \frac{\partial^2}{\partial u(x_i, y_k)^2} F(x_i, y_k),$$

$$\frac{\partial}{\partial t} \bar{F} \quad \text{wie} \quad h_i h_k \frac{\partial}{\partial t} F(x_i, y_k) \quad \text{und}$$

$$\frac{\partial}{\partial \bar{u}} \bar{F} \quad \text{wie} \quad \frac{\partial}{\partial u(x_i, y_k)} F(x_i, y_k) \quad \text{mit } (x_i, y_k) \in \Omega_{i,k}.$$

Geht man davon aus, daß dies unabhängig von der Größe von Ω_{ik} gilt, so läßt sich ein Zusammenhang zwischen dem lokalen Diskretisierungsfehler τ und der Größe $\bar{F}$ ersehen. Terme, die bezüglich Änderungen von Ω_{ik} aufgrund obiger Näherungen konstant sind, werden in folgender Darstellung nicht mehr explizit berücksichtigt:

$$4.8.2 \quad \tau_{\Delta t}(t) = \Delta t^2 h_i h_k F(x_i, y_k) \text{ Constant.}$$

Der zeitliche Diskretisierungsfehler wird durch die Kantenlängen h_i , h_k und die $F(x_i, y_k)$ bestimmt. Wird das Gitter so gelegt, daß $h_i h_k F(x_i, y_k)$ für alle i, k gleichverteilt ist, so wird nach (4.8.1) auch $\tau_{\Delta t}$ im Schnitt minimiert. Eine Gleichverteilung der rechten Seite von (4.8.2) hat auch eine Auswirkung auf die "Steifigkeit" des Systems der gewöhnlichen Zeitdifferentialgleichungen.

Ein strenger Nachweis ist dafür nicht möglich, da die rechten Seiten der Gleichungen nichtlinear sind und diese, wie in [GRI-72] gezeigt, zu unverhofften Stabilitätsschwierigkeiten führen können. Zumindest ist eine plausible Darstellung möglich, die den Einfluß von $\bar{F}$ auf die Steifigkeit der Gleichungen zeigt.

Die rechte Seite des Systems der gewöhnlichen Differentialgleichungen soll durch eine konstante Matrix Q_c approximiert werden, die zur besseren Übersicht diagonalähnlich sein soll. $q_1, \dots, q_{n^2}$ bilden in aufsteigender Form die Eigenwerte der Matrix Q_c . Die allgemeine Lösung [WAL-76] des approximierten Systems ergibt sich bekanntlich als Linearkombination der skalaren Differentialgleichung $v' = q_k v$. Ein von Dahlquist [GRI-72] eingeführter Stabilitätsbegriff fordert, daß die Näherungslösung fallen soll, wenn es die zu approximierende Lösung tut. Aus der Diskretisierung der skalaren Differentialgleichung mit der Trapezregel ergibt sich folgende Rekursion für die approximierte Lösung $v^{j+1} := v(j \times \Delta t)$:

$$v^{j+1} = v^j + \frac{1}{2} \left(\frac{1 + \Delta t q_k}{1 - \Delta t q_k} \right) v^j = v^0 + \left(\frac{1 + \Delta t q_k}{1 - \Delta t q_k} \right)^j v^0 + (g_s(\Delta t q_k))^j.$$

Analog ist g_s für das approximierte System $u' = Q_c u$ definiert:

$$4.8.3 \quad u^{j+1} = u^0 + (g_s(\Delta t Q_c))^j.$$

Damit zur Definition von Dahlquist:

4.8.4 Definition

Ein Verfahren heißt absolut stabil, wenn für die Rekursion g_s gilt:

$$|g_s(z)| < 1 \quad \text{für } \operatorname{Re}(z) < 0.$$

Wegen

$$4.8.5 \quad \left| \frac{1+z}{1-z} \right| < 1 \quad \text{für } \operatorname{Re}(z) < 0.$$

ist also die Trapezregel absolut stabil.

Von einem steifen System spricht man, wenn es zwei Eigenwerte q_k und q_l mit $\operatorname{Re}(q_k) < 0$ und $\operatorname{Re}(q_l) < 0$ gibt, für die $\operatorname{Re}(q_k)$ sehr viel kleiner ist als $\operatorname{Re}(q_l)$. Die Matrix Q_c ist diagonaldominant, somit werden die Eigenwerte vorallem durch die Diagonalelemente bestimmt, in die wiederum $h_i h_k F(x_i, y_k)$ eingeht. Für eine gute Konditionierung der Eigenwerte ist also wieder eine Gleichverteilung der $h_i h_k F(x_i, y_k)$ dienlich.

In 4.7 ist eine Strategie entwickelt worden, die die Minimierung der Ortsdiskretisierung zum Inhalt hat und in 4.8 ist die Auswirkung des Gitters auf die Zeitdiskretisierung untersucht worden. Geht man davon aus, daß große Werte von $F(x_i, y_k)$ meist durch hohe zweite Ableitungen des elektrostatischen Potentials bedingt sind, erhält man ähnlich wie in 4.7.1, als Richtbedingung für die Anpassung an die Zeitdiskretisierung:

$$4.8.6 \quad h_i h_k |\Phi''_{ik}| = \text{const}, \quad \forall i, k = 1, \dots, n.$$

Je nachdem, ob eine gute Orts- oder Zeitdiskretisierung bevorzugt wird, ist das Gewicht bei der Gittergeneration auf (4.7.1) oder (4.8.6) zu legen. In den gerechneten Beispielen sind die Ergebnisse dann am besten gewesen, wenn das Hauptgewicht auf (4.8.6) lag. Die Ortsdiskretisierung ist bei nach (4.8.6) generierten Gittern auch sehr gut, was umgekehrt nicht immer der Fall war. Deswegen wird in der Anwendung auf die Einhaltung von (4.8.6) geachtet, und nur wenn möglich (4.7.1) berücksichtigt.

4.9 Konsistente Berechnung der Lösungswerte auf einem neuen Gitter

Ist ein neues Gitter gefunden, müssen die Funktionswerte an den neuen Stützstellen berechnet werden. Da man eine Polynomdarstellung der Lösung auf dem alten Gitter besitzt, ist im Prinzip eine Neuberechnung recht einfach. Jedoch ist eine durch die besondere Art der Ansatzfunktionen bedingte Schwierigkeit zu beachten. Es treten, wie schon oben beschrieben, Unstetigkeiten am Interface auf, die eine eindeutige Darstellung an den Rändern nicht zulassen. Da eine neue Darstellung, beispielsweise bei einer Herausnahme einer Gitterlinie, diese Unstetigkeit nicht mehr haben kann, stellt sich die Frage nach einer konsistenten Darstellung, wobei Konsistenz im Sinne von Massenerhaltung gemeint ist. Die Massenerhaltung ist gegeben, wenn die Summe der Integrale der Lösung über die ASWR-Boxen (vgl. (2.1.1)) vor und nach Gitterveränderung gleich bleibt. Damit läßt sich ein konsistentes Verfahren entwickeln. Die Berechnung der neuen Werte geschieht über die Integrale. So werden zuerst die neuen Integrale auf dem neuen Gebiet mit Hilfe der alten Darstellung der Lösung berechnet, danach werden mit Hilfe eines Gleichungssystems die B-Spline-Koeffizienten auf dem neuen Gitter bestimmt. Die Darstellung des Gleichungssystems läßt sich aus Kapitel 2 ersehen.

4.10 Fehlerbetrachtung durch Massenerhaltung

Wie in 4.1 gezeigt, wird die Umverteilung im wesentlichen durch das modifizierte Fick'sche Gesetz bestimmt. Gelten an den Rändern Neumann-Bedingungen, ist also der Massenfluß am Rand gleich Null, so sollte, gemäß dem physikalischen Modell, die Masse während der Diffusion gleichbleiben. Im letzten Absatz ist diese Eigenschaft zur Uminterpolation bereits ausgenützt worden. Dort bleibt die Summe der Integrale über den Boxen bei der Umrechnung gleich. Zeit- und Ortsdiskretisierung sind zusammen nicht unbedingt "massenerhaltend", was bedeutet, daß die Diskretisierung unabhängig von deren Güte Massenerhaltung erzwingt. Es gibt zum behandelten Problem konsistente Differenzenschemata, die für jedes beliebige Gitter, sofern das dazugehörige Gleichungssystem korrekt gelöst wird, massenerhaltend sind. Das gilt im allgemeinen für das ASWR-Verfahren nicht. Hier ist Massenerhaltung mit Sicherheit nur im Rahmen der Güte der Orts- und Zeitdiskretisierung zu erhalten. Damit kann die Massenbilanz für die Güte der ASWR-Simulation verwendet werden. Zwar ist einschränkend die Qualität der Massenbilanz nicht die direkte Folge der Qualität der Diskretisierung, da es auch vor-

stellbar ist, daß Fehler sich im Schnitt kompensieren können. In Anbetracht dessen, daß gerade im nichtlinearen Fall nur sehr grobe Fehlerabschätzungen möglich sind (wenn sie überhaupt möglich sind), ist zumindest die Massenbilanz ein guter und praktikabler Hinweis auf die Simulationsgüte.

5. Durchführung der ASWR-Methode für die Festkörperdiffusion

In diesem abschließenden Kapitel werden zwei numerische Beispiele aus der CMOS-Technologie gezeigt. Ein drittes ist nicht direkt aus der Anwendung entnommen, soll jedoch die physikalischen und numerischen Zusammenhänge noch einmal bildlich darlegen. Alle drei Beispiele beschreiben die gekoppelte Diffusion zweier Dopanden.

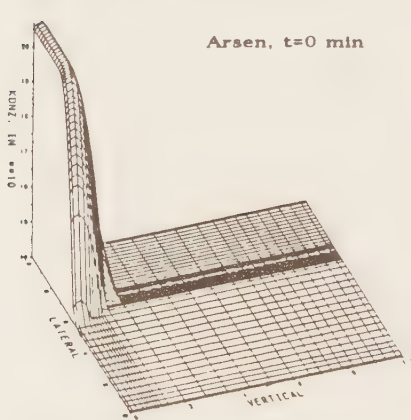


Abbildung 1

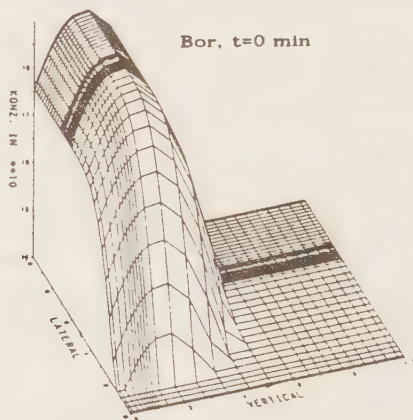


Abbildung 2

5.1 Kopplung von Arsen- und Bordiffusion

Wie in Kapitel 4 erläutert, muß für jeden Dopanden jeweils eine eigene partielle Differentialgleichung berechnet werden, welche über die Diffusionskoeffizienten und das elektrostatische Potential gekoppelt sind. Interessant ist vor allem die Interaktion zwischen der Donatoren- und Akzeptorengleichung, die ja entgegengesetztes elektrisches Verhalten zeigen. Durch Donatorenkonzentrationen wird die Elektronendichte erhöht und zugleich die "Löcherkonzentration" erniedrigt, bei Akzeptorenkonzentrationen ist der Sachverhalt qualitativ genau umgekehrt. Wenn auch das nächste Beispiel nicht direkt einer Anwendung entnommen ist, so sollen damit die Zusammenhänge nichtintrinsic Diffusion verdeutlicht werden. Außerdem demonstriert das Beispiel die numerische Leistungsfähigkeit des ASWR-Verfahrens; wegen der hohen also nichtintrinsic Konzentration sind hier Nichtlinearitäten, Kopplung der Gleichungen, Steifigkeit der Zeitgleichungen, schneller Wechsel von hohen Gradienten und damit verbundene Anforderungen an das Rechengitter besonders deutlich ausgeprägt. Als Donatorenkonzentration wurde eine Arsenimplantation mit einer

Spitzenkonzentration von $5 \times 10^{20} \text{ cm}^{-3}$, einer Durchschnittsimplantationstiefe von $.025 \mu\text{m}$ und einer Streuung von $0.025 \mu\text{m}$ gewählt. Die Konzentrationskurve wird mit einer Gaussglocke beschrieben, wobei die Durchschnittsimplantations-tiefe der Mittelwert der Kurve ist [Abb.1.]. Der Akzeptorendopant Bor hat eine Spitze.konzentration von 10^{19} , eine Durchschnittsimplantationstiefe von $.18 \mu\text{m}$ und eine Streuung von $.07 \mu\text{m}$ [Abb.2.].

5.1.1 Bemerkung zu den Plots

Die Plots repräsentieren ASWR-Lösungen nur in den Schnitten der Gitterlinien. Die Schnitte bezeichnen die Mittelpunktswerte in dem ASWR-Gitter. Somit stimmen die Gitterlinien im Plot nicht überein mit dem verwendeten ASWR-Gitter, geben aber trotzdem einen guten Überblick über das benutzte Rechengitter, da ja im Inneren des Gebiets zwischen zwei Plotlinien eine ASWR-Gitterlinie liegen muß. Die Interpolation zwischen den Gitterlinien erfolgt nur linear, obwohl laut ASWR-Ansatz eine quadratische Funktion die Lösung beschreibt. Die Plot-Software fügt jedoch bei quadratischer Interpolation "Effekte" hinzu, die tatsächlich nicht berechnet werden.

5.1.2 Gittergeneration

Als Eingabe für dieses Beispiel erhält das Programm ein äquidistantes Gitter. Vor dem ersten Zeitschritt wird nun das Gitter nach der in Kapitel 4 beschriebenen Strategie umsortiert. Wie in Abbildung 3 zu sehen ist, richtet sich der Gittergenerator in erster Linie nach dem p-n-Übergang, das heißt nach dem steilen Abfall im elektrostatischen Potential. Dabei wurden, obwohl dies

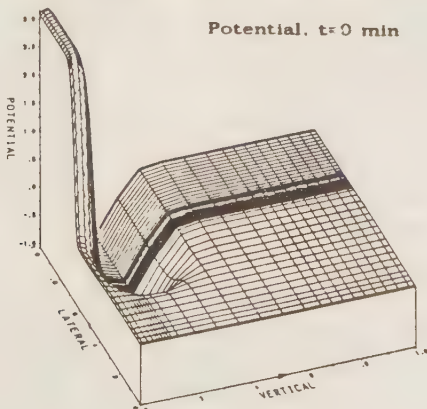


Abbildung 3

nicht notwendig wäre, in lateraler Richtung 50 Gitterlinien verwendet, um die Wirkung des Gittergenerators zu demonstrieren. Die vertikale Richtung wurde mit 25 Gitterlinien versehen; verglichen dazu hätten in lateraler Richtung 20 Gitterlinien in etwa ausgereicht. Der Unterschied des Gitters in lateraler und vertikaler Richtung erlaubt einen Vergleich der Generierung mit unterschiedlicher Anzahl von Gitterlinien. Der starke Abfall des elektrostatischen Potentials in lateraler Richtung und die damit verbundene Häufung von Gitterlinien erklärt sich durch die Arsenkonzentration, die aufgrund einer Maske zwischen 0.7μ und 1.0μ implantiert wurde. Die negativen Werte des Potentials werden durch nichtintrinsische Konzentrationen des Bors verursacht, sodaß der Abfall in vertikaler Richtung größer ist als in lateraler und folglich bei gleicher Anzahl der Gitterlinien in beiden Raumrichtungen zu einer eher noch stärkeren Häufung als in lateraler Richtung hätte führen müssen. Die Häufung der Gitterlinien wird, wie man deutlich sehen kann, unter Vermeidung neuer großer Gitterabstände erzeugt. Das minimiert bei wandernden Profilen, deren Richtung man nicht genau kennt, das Umsetzen von Gitterlinien. Dieses Beispiel ist so gewählt, daß der Gittergenerator ungünstigerweise dort Linien herausnimmt, wo er sie während der fortschreitenden Simulation wieder einfügen muß. Trotzdem hält sich die Anzahl der Umsortierungen in Grenzen.

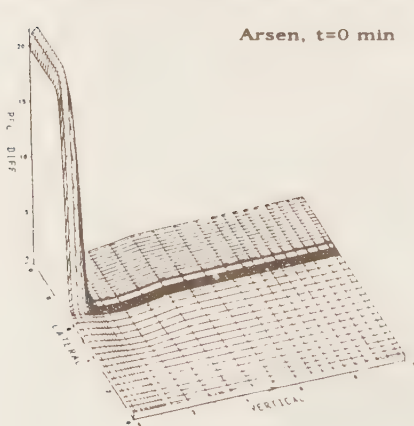


Abbildung 4

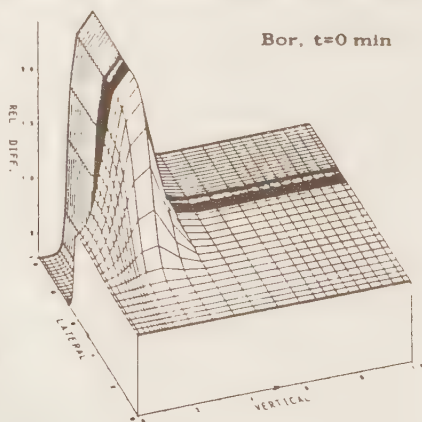


Abbildung 5

In Abbildung 4 und 5 sind die Auswirkungen der Dopanten auf die relativen Diffusionskoeffizienten (relativ zum intrinsischen Diffusionskoeffizienten) zu sehen. Der relative Diffusionskoeffizient von Arsen ist natürlich bei hoher Konzentration von Arsen sehr groß, fällt jedoch unter den Wert 1 bei hoher Konzentration von Bor, was bedeutet, daß der Diffusionskoeffizient unter

dem des intrinsischen liegt. Gleiches Verhalten, jedoch nur mit umgekehrten Vorzeichen, ist auch beim relativen Borkoeffizienten zu beobachten. Dort wird der Diffusionskoeffizient bei hohen Arsenkonzentrationen unter den des intrinsischen gedrückt. Relativer Diffusionskoeffizient und elektrostatisches Potential sind beide in erster Linie von den Elektronen- bzw. Löcherkonzentrationen abhängig, das Gitter ist somit auch dem Verlauf des relativen Diffusionskoeffizienten gut angepaßt.

5.1.3 Die Anpassung des Gitters an die physikalischen Erfordernisse im zeitlichen Verlauf

Bis jetzt erfolgte die Anpassung des Gitters an ein statisches Problem, nämlich der Implantationsverteilung. Mit der Ausheilung der durch die Implantation gestörten Kristallgitter wird zugleich eine Diffusion der implantierten Fremdatomkonzentrationen erreicht. Die Profile wachsen auseinander; es kommt jedoch auch zu Kopplungseffekten, die vom Gitter entsprechend gewürdigt werden müssen. Abbildung 6, 7, 8 und 9 zeigen das Bor-Profil nach 1, b.z.w. 5 b.z.w. 10 b.z.w. 30 Minuten Diffusionszeit bei 950° Grad (Temperaturangaben verstehen sich immer in Celsius). Dabei ist deutlich das Wachsen eines

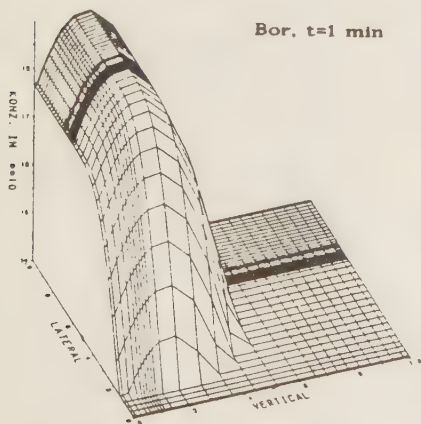


Abbildung 6

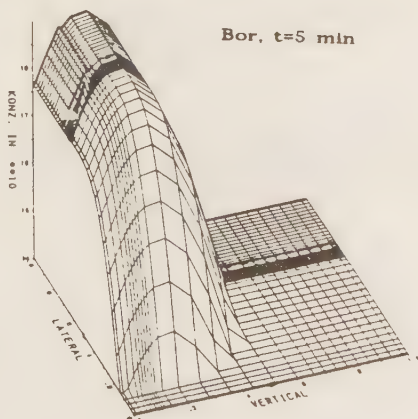


Abbildung 7

Sattels zu beobachten. Durch den Einfluß des hohen Gradienten des elektrostatischen Potentials bildet das Borprofil ein neues lokales Maximum. Dem Fluß hoher n-Dotierungen wird im Bemühen, diese zu neutralisieren, ein Fluß von p-Dotierungen entgegengehalten. Dies bewirkt das neue lokale

Maximum, zugleich aber auch einen Graben in der Bordotierung um dieses neue Maximum. Zu sehen ist dieser Graben auch in lateraler Richtung am lateralen Ende des neuen lokalen Maximums. Dieser Effekt ist somit nicht nur vertikaler Natur. Nicht ganz zufällig ist schon vor Ausbildung des Effekts das Gitter an den signifikanten Stellen feiner. Der Gittergenerator richtet sich, wie oben schon erwähnt, nach dem elektrostatischen Potential, häuft somit in den Potentialflüssen die Gitterlinien. Die Potentialflüsse sind aber auch für obigen Effekt verantwortlich; somit ist die Gitterstruktur schon vor Auftauchen des Effektes günstig für eine genaue Darstellung, was einem impliziten Gittergenerator, also einem schon für den nächsten noch nicht berechneten Zeitschritt optimalen Gittererzeuger, entspricht. Erst nach 30 Minuten wird das Gitter an besagter Stelle in lateraler Richtung noch einmal verfeinert [Abb.9]. Bis dahin wurde das Gitter in lateraler Richtung nicht

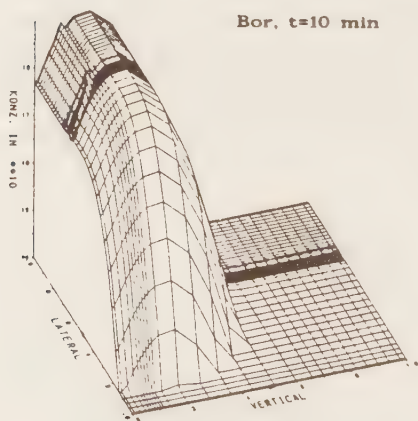


Abbildung 8

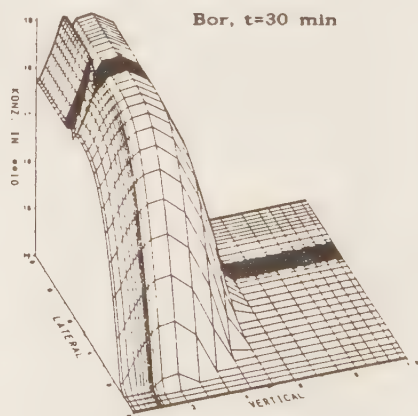


Abbildung 9

verändert. Daß der Effekt nur durch das elektrostatische Potential erreicht wird, sieht man in Abb. 10 am relativen Diffusionskoeffizienten von Bor, der dort unter 1 ist; der Diffusionskoeffizient also ist dort kleiner als der intrinsische. Zudem differiert er lokal nicht, was einen solchen Effekt erklären ließe. Die neu eingefügten Linien tragen dem steileren Gradientenverlauf des elektrostatischen Potentials Rechnung. Das Arsen diffundiert in der Zwischenzeit am Rand aus, da nur hier die Diffusionskoeffizienten entsprechend hoch sind. Der zum Gebietsinneren steiler werdende Abfall, bedingt durch die stark unterschiedlichen Diffusionskoeffizienten [Abb.11,12] läßt auch das elektrostatische Potential steiler werden [Abb.13] und bewirkt damit die Einfügung der Gitterlinien.

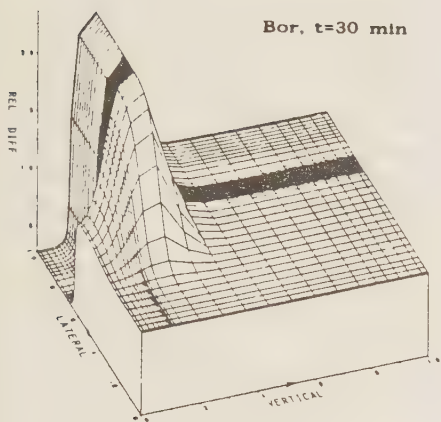


Abbildung 10

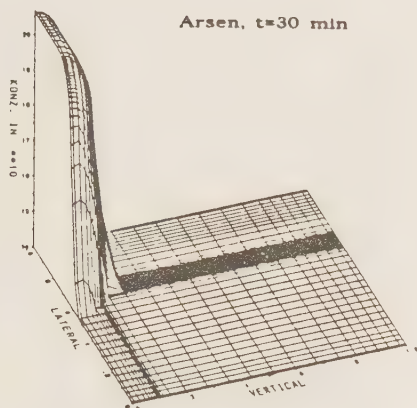


Abbildung 11

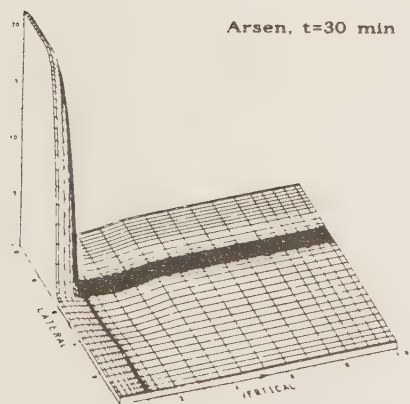


Abbildung 12

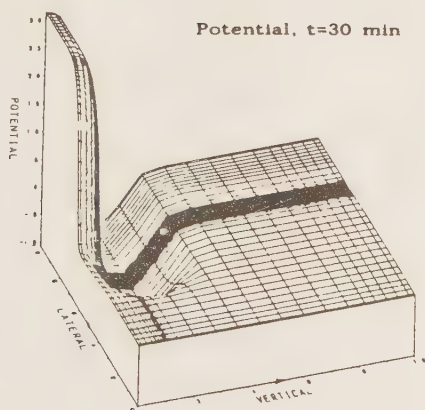


Abbildung 13

5.2 Simulation einer CMOS-Struktur

Das Programm ist für die Simulation bestimmter Fabrikationsprozesse gedacht. So wird nach einer Implantation das Kristallgitter des Siliciums gestört, was auch die elektrischen Eigenschaften des Siliciums verändert. Eine Möglichkeit, die ursprünglichen Eigenschaften des Siliciums zu restaurieren, ist das Ausheilen des dotierten Siliciums unter Einwirkung hoher Temperaturen. Dabei verändern sich auch die Dotierungsprofile; Diffusion bewirkt ein Auseinanderlaufen der Profile. Es können jedoch auch Effekte auftreten wie im obigen Beispiel. Die Lage des p-n-Übergangs kann sich entscheidend verändern, sodaß im schlechten Falle Kurzschlüsse oder andere nichtgewollte Effekte auftreten können. Um solches zu vermeiden oder genauere Kenntnisse über bestimmte Zusammenhänge zu erfahren, werden die Prozeßschritte simuliert, da eine Testreihe wesentlich mehr Aufwand erfordert. Im folgenden soll die Dopandenverteilung nach der Ausheilung einer Implantation im Drain/Source-Gebiet und einer Kanalimplantation in einem p- wie auch n-Gebiet berechnet werden. In Abbildung 14 ist die gemeinsame Anordnung eines p- und n-Kanals zu sehen. Simuliert wird jeweils getrennt die Entstehung eines p- und n-Kanals. Zunächst die Simulation im n-Gebiet.

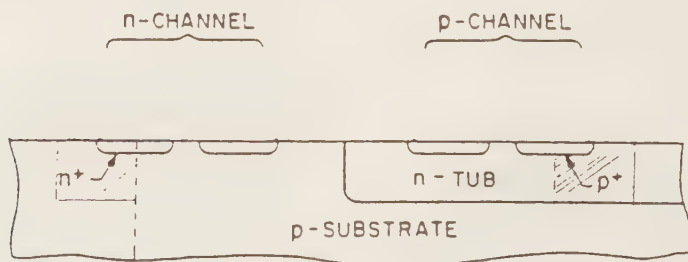


Abbildung 14

5.2.1 Simulation im n-Gebiet

In der rechten Hälfte der Abbildung 14 ist der p-Kanal zu sehen. Mit Hilfe einer Maske erfolgt die Drain/Source-Implantation. Die Maske vermeidet eine Dopandenimplantation an der Stelle des Gates. Um den Rechenaufwand

zu verkleinern, wird nur auf einem Teilgebiet gerechnet. Man geht davon aus, daß aus Symmetriegründen die linke Hälfte (getrennt durch die gestrichelte Linie) sich genauso verhält wie die rechte. Als Randbedingung für diesen "künstlichen" Rand wird Fluß gleich 0 angenommen, ebenso wie an den drei anderen Rändern. Um eine Grunddotierung mit n^- zu erhalten, wird Phosphor in das Gebiet implantiert und mit 1100° ausdiffundiert, bis man eine Gleichverteilung erhält. Das Niveau liegt knapp unter 10^{17} cm^{-3} , die Ausdiffusion erfolgt 900 Minuten. Für die Drain/Source-Implantation von Bor mit einer Energie von 50 keV und einer Dosis von 10^{16} cm^{-3} wird eine Maske verwendet, die 0.5 μ des Rechengebietes, der den Kanalbereich darstellt, abdeckt. 30 Minuten Gattern bei einer Temperatur von 900° sollen das dotierte Silicium von störenden Atomen befreien, das 20 minütige Ausheilen bei 950° beendet den Simulationsprozeß.

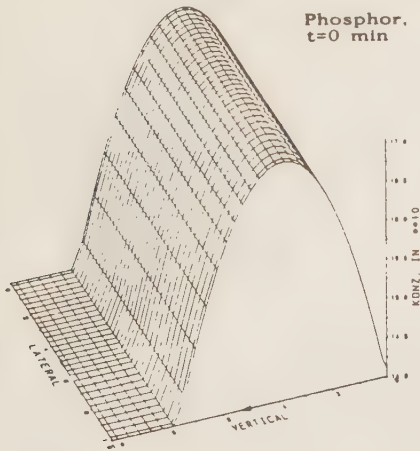


Abbildung 15

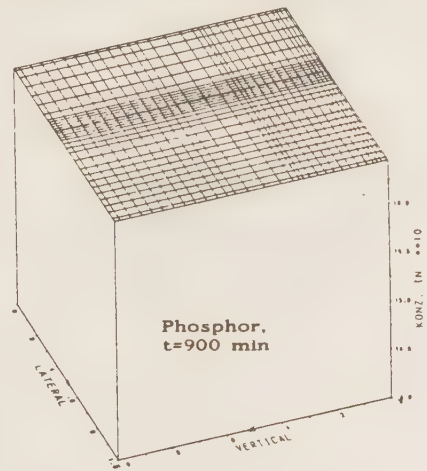


Abbildung 16

Zur besseren Übersicht werden folgende Abbildungen aus einem anderen Blickwinkel betrachtet (Drehung um 180°). Abbildung 15 zeigt das Phosphorprofil nach der Implantation und Abbildung 16 nach der Ausdiffusion. Während der Ausdiffusion wurde ein äquidistantes Gitter verwendet; in Abbildung 16 ist das Gitter schon an die Borimplantation angepaßt. Das Borprofil ist in Abbildung 17 zu sehen. Abbildung 18 stellt das Borprofil nach 30 Minuten Diffusion bei 900° dar. Durch die Nichtlinearitäten ist der Konzentrationsabfall steiler geworden; das Gitter hat sich der Situation angepaßt und ist am Abfall dichter. Abbildung 20 zeigt das elektrostatische Potential, das ja die Gittererzeugung im wesentlichen bestimmt. Außerdem bewirkt das Borprofil in der vor der Borimplantation gleichverteilten Phosphorkonzentration

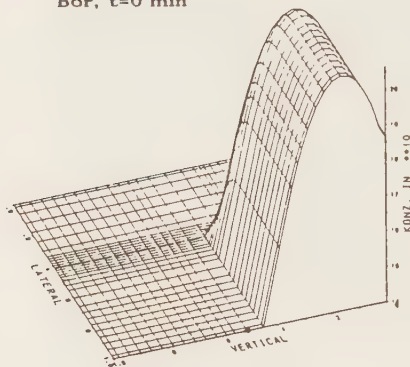
Bor, $t=0$ min

Abbildung 17

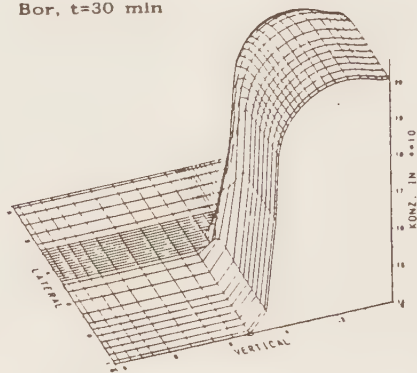
Bor, $t=30$ min

Abbildung 18

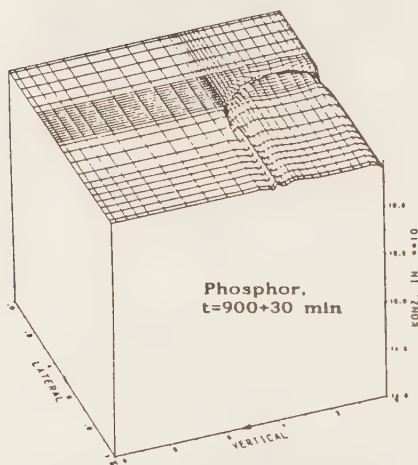


Abbildung 19

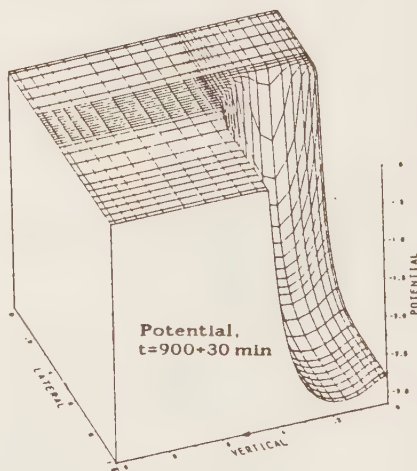


Abbildung 20

einen Graben mit einem spitzen Maximum. Dies wird durch hohe elektrostatische Potentialgradienten, die wiederum durch die hohen Borgradienten bedingt sind, erwirkt. Somit können also auch die Donatoren Auswirkungen auf die Akzeptorenkonzentrationen haben, so wie umgekehrt im vorherigen

Beispiel. Nach weiteren 50 Minuten Diffusion bei 900° ist das Borprofil [Abb.21] wieder ein relativ großes Stück weiterdiffundiert und hat das Gitter mit Hilfe der Potentialverteilung "mitgenommen" [Abb.23]. Abbildung 22 zeigt im Phosphorprofil den mittlerweile verbreiterten Graben. Die im vorherigen Prozeßschritt noch beobachteten lokalen Maxima sind auseinanderdiffundiert und nicht mehr zu sehen.

Bor, $t=80$ min

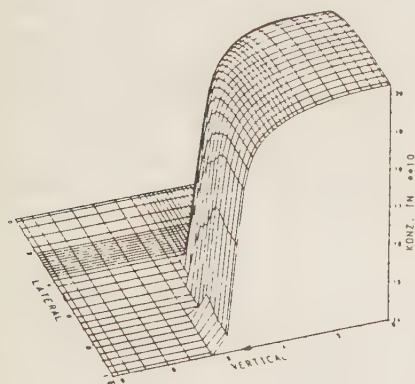


Abbildung 21

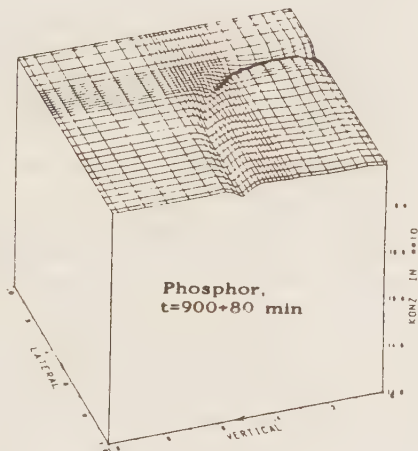
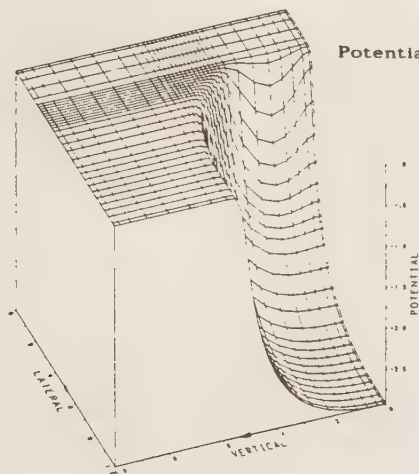


Abbildung 22



Potential, $t=900+80$ min

Abbildung 23

5.2.2. Simulation im p-Gebiet

Typisch für die Architektur der CMOS-Struktur ist das Vorhandensein von p-Wannen, also Dotierverteilungen "umgekehrten Vorzeichens" zum letzten Beispiel. Hier ist die Grunddotierung mit p^+ vorbesetzt und die Drain/Source-Implantierung erfolgt mit n^- geladenen Stoffen. Genau wie oben wird hier die Symmetrie des Gebietes ausgenützt. Wurde letztes Beispiel mit einem Gitter aus 30×30 Linien gerechnet, so werden hier nur 20×20 verwendet. Da der Schwierigkeitsgrad des letzten und des jetzigen Beispiels in etwa gleich sind, kann man somit die Güte der Lösung bei abnehmender Zahl der Gitterlinien vergleichen. Es stellt sich dabei heraus, daß keine signifikanten Unterschiede festgestellt werden können. Auch die Massenbilanz, die ja als Gütemerkmal verwendet werden kann, unterscheidet sich in beiden Fällen um nicht mehr als zwei Prozent. Daraus kann man schließen, daß Gitter von 20×20 Linien für diese Beispiele voll ausreichend sind. Wie im jetzigen Beispiel noch gezeigt wird, sind damit durchaus "feine Effekte", wie die einseitige Auswirkung der Arsendotierung auf eine niedrige Borverteilung zu beobachten.

Um eine p^+ -Grunddotierung zu erhalten, wird eine breit gestreute Borimplantation vorgenommen. Nach 6 Stunden bei 1100° ist das Profil ausdiffundiert, sodaß man eine Gleichverteilung des Bors bei etwa $6 \times 10^{16} \text{cm}^{-3}$ erhält [Abb. 24], [Abb. 25]. Daraufhin erfolgt eine Arsenimplantation mit 120 keV und einer

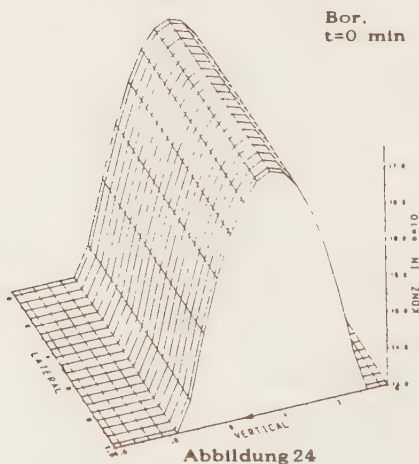


Abbildung 24

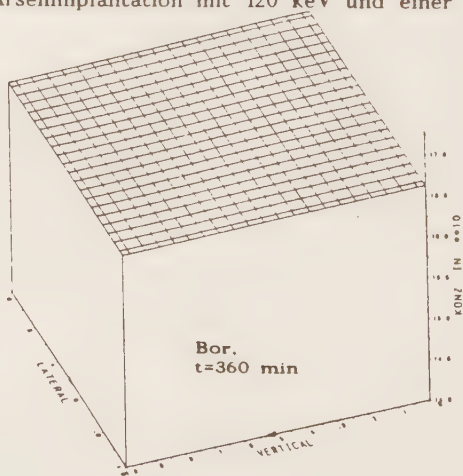


Abbildung 25

Dosis von $5 \times 10^{14} \text{cm}^{-2}$. Die Spitzenkonzentration liegt knapp unter 10^{20}cm^{-3} [Abb. 26]. War für die Ausdiffusion des Borprofils das Gitter wegen der großen Streuung noch äquidistant gewesen, wird das Gitter nun deutlich

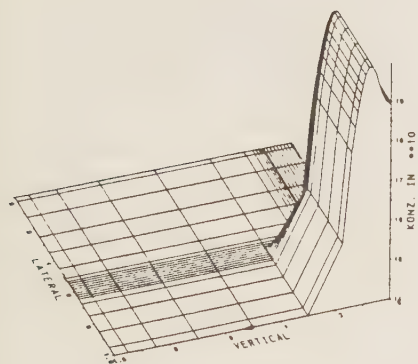
Arsen, $t=0$ min


Abbildung 26

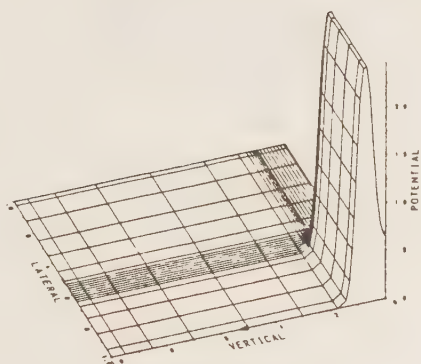
Potential, $t=360+0$ min


Abbildung 27

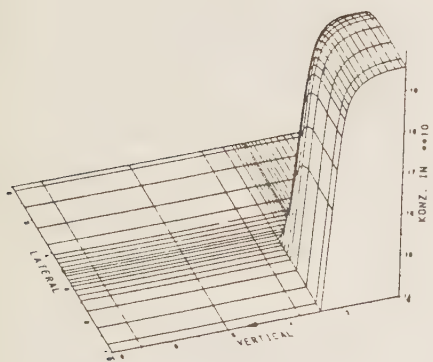
Arsen, $t=40$ min


Abbildung 28

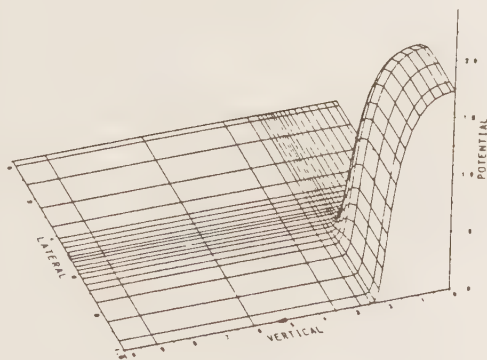
Potential, $t=360+40$ min

Abbildung 29

anders verteilt. In Abbildung 27 sieht man die Potentialverteilung, die Grundlage für das neue Gitter ist. Nach 60 Minuten Diffusion bei 1000° ist das Arsenprofil deutlich auseinandergeschoben [Abb. 28]. Das Mitwachsen des Gitters mit dem Profil ist beim Vergleich der Abbildungen 27 und 29 gut zu erkennen; es ist eine größere Struktur des Gitters nach 60 Minuten zu beobachten, bedingt durch den breiteren Profilverlauf der Potentialverteilung. Die Auswir-

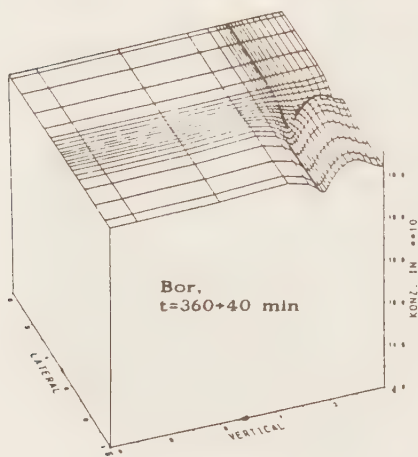


Abbildung 30

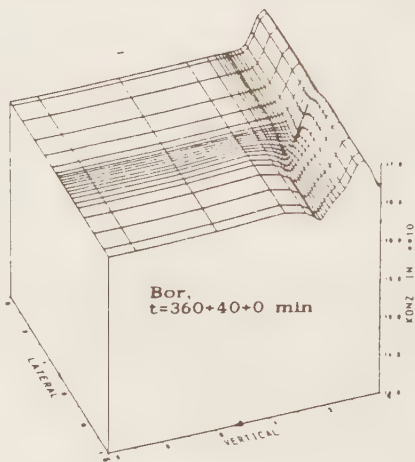


Abbildung 31

Arsen, $t=40+20$ min

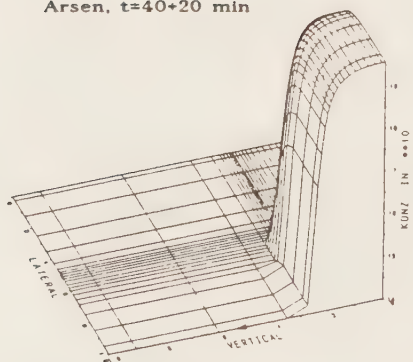


Abbildung 32

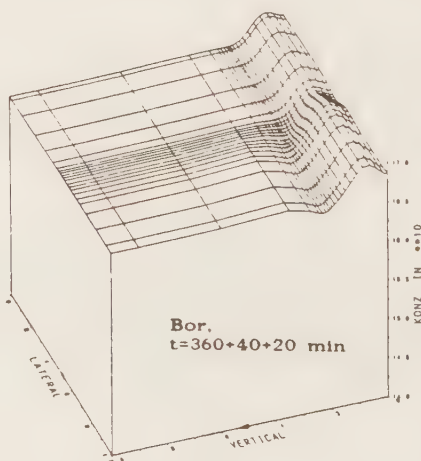


Abbildung 33

kungen des Arsenprofils auf das ehemals gleichverteilte Bor ist in Abbildung 30 zu sehen. Dabei sind wie in 5.1 ein Graben und ein lokales Maximum zu entdecken. Es erfolgt nun als weiterer Prozeßschritt eine Kanalimplantation von Bor mit einer Dosis $5 \times 10^{11} \text{cm}^{-3}$ bei 25 keV. In Abbildung 31 sieht man dem neuen Borprofil den Effekt aus Abbildung 30 noch an. Nach weiteren 20 Minuten Diffusion bei 950° sind die Änderungen im Arsenprofil gering, da

wegen der niedrigeren Prozeßtemperatur die Diffusionskoeffizienten wesentlich kleiner sind [Abb. 32]. Im Bor-Profil ist weiterhin der Einfluß der Arsendotierung sichtbar. Während an den Stellen, an denen auch die Arsenkonzentrationen sehr hoch sind, die Bordiffusion gebremst erscheint, ist diese andererseits weiter vorgeschritten. Das führt auf den auch mit Arsen dotierten Gebieten zu weniger breiten, dafür noch höheren Borkonzentrationen, andererseits sind die Borverteilungen breiter und niedriger. Eine Mulde in vertikaler Richtung ist weiterhin an der Implantationswand zu sehen, genauso eine Mulde in lateraler Richtung am Bergende, da ja das Arsen auch weiterhin auf das Borprofil wirkt. Im letzten Prozeßschritt ist der Gittergenerator nicht mehr in Kraft getreten, sodaß dieser nur während der 60 Minuten Diffusionszeit bei 10000° aktiv sein mußte.

5.3 Ausblicke

Die Leistungsfähigkeit des Programms liegt natürlich nicht nur in der genauen Simulation, sondern auch in der effizienten Anzahl und Generierung der Gitterlinien und der damit verbundenen geringen Rechenzeit. So konnte, obwohl das Programm bezüglich der Rechengeschwindigkeit noch nicht optimiert ist, ein Speedup-Faktor gegenüber dem alten Diffusionsmodul [SVO-85] von 10 und mehr erreicht werden. Zwar ist der Vergleich zweier Programme gerade bei solcher Komplexität sehr schwierig, jedoch ist zumindest für die bis jetzt gerechneten Beispiele ein deutlicher Vorteil in Bezug auf Rechengenauigkeit und Rechengeschwindigkeit gegenüber dem alten Programm zu ersehen. Die Simulation von Mehrschichtdiffusion, die wegen der zu erwartenden hohen Rechenzeiten mit dem alten Modul [SVO-85] nicht realisiert worden ist, wird mit Hilfe des ASWR-Verfahrens angegangen werden. Auch die Anwendung auf dreidimensionale Raumstrukturen ist durchaus vertretbar.

Die Hauptlast der Rechenzeit liegt in der Berechnung der B-Spline-Koeffizienten und der Stützstellenwerte. Man kann davon ausgehen, daß dafür mindestens 90% der Rechenzeit angesetzt werden muß. Sei es die Berechnung der B-Spline-Koeffizienten durch Lösen des Gleichungssystems mit Hilfe des ADI-Verfahrens oder die Berechnung der Stützstellenwerte, beide Operationen lassen sich einfach und effektiv vektorisieren. Hier sind sicherlich noch große Möglichkeiten der Rechenzeitoptimierung offen, die, je nach Rechenmaschine, dreidimensionale Probleme in die Rechenzeiten zweidimensionaler oder sogar darunter führen können.

Literatur

- [BAK-76] G. A. Baker : Error estimates for Finite Element for second-order hyperbolic equations,
SIAM J. Numer. Anal. 13, (1976) S. 564-576
- [BOO-78] C. de Boor : A Practical Guide to Splines,
Springer Verlag, New York (1978)
- [DUT-82] D. Chin, M. R. Kump, H. G. Lee, R. W. Dutton : Process Design using two dimensional process and device simulators,
IEEE Trans. Electron Devices, vol. ED-27, S. 336 (1982)
- [GRA-78] Graf : Persönliche Mitteilungen,
Gesellschaft für Reaktorsicherheit (1978)
- [GRO-66] Groebner, W. : Matrizenrechnung,
B. I., Mannheim (1966)
- [GRI-72] R.D. Grigorieff : Numerik gewöhnlicher Differentialgleichungen
Band 1 & 2 B. G. Teubner, Stuttgart (1972)
- [HÄM-70] G. Hämmerlin : Numerische Mathematik 1,
BI Hochschultaschenbuch, Mannheim (1970)
- [LAD-68] O. A. Ladyženskaja, V. A. Solonnikov, N. N. Ural'ceva :
Linear and Quasi-Linear Equations of Parabolic Type
American Mathematical Society,
Providence, Rhode Island (1968)
- [LAN-77] S. Langenbuch, W. Maurer, W. Werner : Coarse-Mesh Flux-expansion for the Analysis of Space-Time Effects in Large Light Water Reactor Cores,
Nuclear Science and Engineering 63(1977), S. 437 - 456
- [MAU-77] S. Langenbuch, W. Maurer, W. Werner : High Order Kinetics Calculations, Based on a local Polynomial Approximation,
Nuclear Science and Engineering 64(1977), S. 508 - 516
- [MAL-83] C. D. Maldonado : ROMANS II, a two-dimensional process simulator for modeling and simulation in the design of VLSI devices,
Appl. Phys., vol. A31, S. 119 (1983)

- [MAU-78] W. Maurer : Ein numerisches Rechenmodell für die effiziente Berechnung der ortsabhängigen Leistungsverteilung in großen Leichtwasserreaktoren,
Interner Bericht, Gesellschaft für Reaktorsicherheit (1978)
- [ODE-76] J. T. Oden, J. N. Reddy : An Introduction to the Mathematical Theory of Finite Elements,
Wiley, New York (1976)
- [PER-75] V. Pereyra, E. G. Sewell : Mesh selection for discrete solution in ordinary differential equations,
Numer. Math. 23 (1975), S. 261 - 268
- [PIC-85] P. Pichler : Numerische Simulation kritischer Prozeßschritte in der Halbleitertechnik,
Dissertation, Technische Universität Wien (1985)
- [PEN-83] B. R. Penumalli : A comprehensive two-dimensional VLSI process simulation program, BICEPS,
IEEE Trans. Electron Devices, vol. ED-30, S. 983 (1983)
- [RYS-86] H. Ryssel, H. Betz, J. Lorenz, J. Pelka, W. Krüger, P. Müller, G. Prinke, A. Sachs, M. Svoboda : Zweidimensionale Prozeßsimulation vollständiger technologischer Prozeßabläufe,
5. Zwischenbericht zum Forschungsvorhaben NT 268613,
Erlangen, Berlin (1986)
- [SCH-82] F. Schmerbeck : Zur Konvergenztheorie des ASWR-Verfahrens.
Diplomarbeit, Mathematisches Institut der Ludwig-Maximilians Universität München, München (1982)
- [SCH-80] H. R. Schwarz : Methode der finiten Elemente.
Teubner Studienbücher, Stuttgart (1980)
- [SMI-70] G. D. Smith : Numerische Lösung von partiellen Differentialgleichungen, Vieweg Verlag, Braunschweig (1970)
- [SVO-85] J. Lorenz, J. Pelka, H. Ryssel, A. Sachs, A. Seidl, M. Svoboda : COMPOSITE, A Complete Modeling Program of Silicon Technology,
IEEE computer-aided design of integrated circuits and systems, S. 421 (1985)

- [SVO-83] M. Svoboda : Das ASWR-Verfahren für zeitabhängige insbesondere hyperbolische Probleme,
Diplomarbeit, Mathematisches Institut der Ludwig-Maximilians-Universität München, München (1983)
- [STO-73] J. Stoer, R. Bulirsch : Einführung in die numerische Mathematik II, Springer, Berlin (1973)
- [TIE-80] R. Tielert : Two-dimensional numerical simulation of impurity redistribution in VLSI processes,
IEEE Trans. Electron Devices, vol. ED-27, S. 1479 (1980)
- [VEL-76] W. Velte : Direkte Methoden der Variationsrechnung
Teubner Taschenbuch, Stuttgart (1976)
- [WAL-76] W. Walter : Gewöhnliche Differentialgleichungen
Springer-Verlag, Berlin (1976)
- [WER-77] W. Werner : Kinetics of Nuclear Systems (Solution Methods Space-Time Dependent Neutron Diffusion Equation),
Advances in Nuclear-Science and Technology 10 (1977),
S. 313 - 364

Bezeichnungen		Seite
$B_{t,i}(x)$	quadratischer B-Spline zur Partition t	6
$B_{t,i}^D(x)$	normierter quadratischer B-Spline zur Partition t	8
$B_{ik}(x,y)$	zweidimensionaler B-Spline erster Art	20
$\bar{B}_n$	Raum der Produkt-Splines	31
C	Raum der n -stetigen bzw m -stetigen Funktionen in x -Richtung b.z.w. y -Richtung	31
C^*	$C^{2,3} \cap C^{3,2}$ mit Norm $\ \cdot \ _*$	33
C_n	Raum der j -mal stückweise stetigen Funktionen	29
$C(I,X)$	Menge der im Zeitverlauf I stetigen Funktionen	40
$D_{ik}(x,y)$	zweidimensionaler B-Spline zweiter Art	23
D_n	Raum der zweidimensionalen Ansatzfunktionen zweiter Art	20
G, G_x, G_y	Operator : Greenscher Operator	13,31
$K_{x,i}(x)$	stückweise Konstante zur Partition x	7
$K_{x,i}^D(x)$	normierte stückweise Konstante zur Partition x	8
K_n	Raum der zweidimensionalen stückweise Konstanten	29
L	Operator : Laplace-Operator	12
L	Operator : zweidimensionaler Laplace-Operator	28
L_x, L_y, L_z	Operator : erweiterte Laplace-Operatoren	29
$P_{2,x}$	Raum der quadratischen Polynomsplines	5
P	Operator : Projektor in $S_{0,x}$	10
Q	Operator : Quasiinterpolierende	7
$\omega(f;h)$	Stetigkeitsmodul	8
Ω	$\Omega_x \times \Omega_y$	31
Ω_1	$(x_{1-1/2}, x_{1+1/2})$	29
Ω_{ik}	$(x_{i-1/2}, x_{i+1/2}) \times (y_{k-1/2}, y_{k+1/2})$	29
Ω_k	$(y_{k-1/2}, y_{k+1/2})$	29
Ω_x	$[a,b]$ in x -Richtung	31
Ω_y	$[a,b]$ in y -Richtung	31
Ω_z	$\bigcup_{i,k=1 \dots n} \Omega_{ik}$	29
r	zweidimensionale Partition := $t \times t$	18
S_n	Raum der zweidimensionalen Ansatzfunktionen erster Art	20
$S_{2,t}$	Raum der quadratischen B-Splines	6
$S_{0,t}$	Raum der stückweisen konstanten B-Splines	6
t	Partition (monotone Folge auf $[a,b]$) auf $[a,b]$	6
T	Operator: Projektion von S_n auf D_n	36
$W_{2,\infty}$	Sobolev Raum	12
W	Operator : Projektor auf S_n	32
x	streng monotone Folge auf $[a,b]$ (Gitterlinien auf den Boxen)	4
$\bar{x}$	streng monotone Folge auf $[a,b]$ (Mittelpunktwerte zwischen den Gitterlinien)	4
X, Y	Operator : jeweilige eindimensionale ASWR-Lösung auf Ω	32
z	zweidimensionale Unterteilung (Boxen)	18

Lebenslauf von Miloš Svoboda

Herkunft

6. 9. 1952 Geburt in München

Ausbildung

1959 - 1964 Volksschule in München

1964 - 1974 Mathematisch naturwissenschaftliches Gymnasium in München

1974 Abitur

1974 - 1983 Studium der Mathematik an der Ludwig-Maximilian-Universität
München

1983 Diplom in Mathematik

Berufliche Tätigkeiten

1978 - 1983 Operateur am Leibniz Rechenzentrum in München

ab 1983 Wissenschaftlicher Mitarbeiter am Fraunhofer Institut in München
und Erlangen

München, im April 1988

